Ana Ramírez López

Improving independent vector analysis in speech and noise separation tasks

School of Electrical Engineering

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 12.04.2015

Thesis supervisor:

Docent Kalle Palomäki

Thesis advisors:

Prof. Nobutaka Ono

M.Sc. (Tech.) Ulpu Remes

Aalto University
School of Electrical
Engineering

AALTO UNIVERSITY
SCHOOL OF ELECTRICAL ENGINEERING

ABSTRACT OF THE
MASTER'S THESIS

Author: Ana Ramírez López

Title: Improving independent vector analysis in speech and noise separation tasks

Date: 12.04.2015 Language: English Number of pages: 6+46

Department of Signal Processing and Acoustics

Professorship: Speech and language processing Code: S-89

Supervisor: Docent Kalle Palomäki

Advisors: Prof. Nobutaka Ono, M.Sc. (Tech.) Ulpu Remes

Independent vector analysis (IVA) is an efficient multichannel blind source separation method. However, source models conventionally assumed in IVA present some limitations in case of speech and noise separation tasks. Consequently, it is expected that using better source models that overcome these limitations will improve the source separation performance of IVA. In this work, an extension of IVA is proposed, with a new source model more suitable for speech and noise separation tasks. The proposed extended IVA was evaluated in a speech and noise separation task, where it was proven to improve separation performance over baseline IVA. Furthermore, extended IVA was evaluated with several post-filters, aiming to realize an analogous setup to a multichannel Wiener filter (MWF) system. This kind of setup proved to further increase the separation performance of IVA.

Keywords: Independent vector analysis, Blind source separation, Microphone array, Speech source model, Speech enhancement

Preface

First of all, I would like to thank my supervisor Docent Kalle Palomäki for his support and guidance in the process of doing this thesis, and also for giving me the opportunity of being part of the Speech Recognition research group of Aalto University. I would also like to thank my two instructors: Professor Nobutaka Ono from the National Institute of Informatics (NII) in Tokyo, Japan, for giving me the opportunity of working on this interesting research topic, and also for his guidance on the theoretical depths of the thesis; Ulpu Remes for her generous effort on guiding me through this thesis and her academic advice.

Also, I want to thank all my colleagues in the Speech Recognition research group, led by Professor Mikko Kurimo, who have always provided a welcoming environment to work in.

Finally, special thanks to my family and friends for being there, and Sergio for his patience and support.

Espoo, 12.04.2015

Ana Ramírez López

Contents

Symbols and abbreviations

m	source channel
M	number of observed signals
N	number of source signals
$S_{m\tau\omega}$	mth source signal (STFT)
$\hat{S}_{m\tau\omega}$	estimate of mth source from single-channel source separation (STFT)
T	total number of time frames
W_ω	demixing matrix
$\boldsymbol{w}_{m\omega}$	mth row of matrix W_ω
$X_{m\tau\omega}$	mth observed signal (STFT)
$\tilde{\boldsymbol{Y}}_{m\tau}$	source-wise vector of mth separated source signal (STFT)
$Y_{m\tau\omega}$	mth separated source signal, and ωth element of $\tilde{\boldsymbol{Y}}_{m\tau}$ vector (STFT)
$\sigma^2_{m\tau\omega}$	variance of extended IVA's statistical source model
τ	time frame
ω	frequency channel

ABF	Adaptative beamformer
BSS	Blind source separation
DFT	Discrete Fourier transform
fwSNRseg	frequency-weighted segmental SNR
ICA	Independent component analysis
IVA	Independent vector analysis
KL	Kullback-Leibler
LTI	Linear time-invariant
MI	Mutual information
ML	Maximum likelihood
MVDR	Minimum variance distortionless response
MWF	Multichannel Wiener filter
NMF	Non-negative matrix factorization
p.d.f	Probability density function
REPET	REpeating Pattern Extraction Technique
RIR	Room impulse response
SISEC	Signal Separation Evaluation Campaign
SDR_i	Signal to distortion ratio
SDR	Source to distortion ratio
SNR	Signal to noise ratio
STFT	Short-time Fourier transform

1 Introduction

In everyday situations, we often encounter noisy environments in which speech perception turns to be a difficult task. This is specially true for older adults, hearing-impaired listeners [1] and children with learning disorders [2]. Past research has proven the importance of social interaction in well-being [3], communication being a key part in social relations. In oral communication, it is essential that the listener perceives and understands correctly the speech signal that is being uttered. As a result of the aforementioned, plenty of research has been done on speech enhancement methods, which aim to improve the quality or intelligibility of speech signals degraded by noise [4, 5].

The situations in which the speech signal of interest is disturbed by competing voices is commonly called the *cocktail party problem* [6], which was formulated by [7] as: "How do we recognize what a person is saying when others are speaking at the same time?". Solving the cocktail party problem is quite complex for man-made systems, like automatic speech recognizers (ASR), while in contrast listeners with normal hearing are able to extract the desired speech signal and neglect the interfering sources with small effort. How the human auditory system solves the cocktail party problem takes advantage of human binaural hearing, and therefore, of the acoustic information gathered at our two ears. The benefits of binaural hearing leads us to source separation algorithms that use multiple microphones, i.e. a *microphone array*, as a promising approach to solve the cocktail party problem [8, 9]. Source separation approaches using multiple microphones are called *multichannel source separation methods*, in contrast to *single-channel source separation*, which employs only one microphone.

Multichannel source separation methods can be classified into two main categories: beamforming and blind source separation (BSS). Beamforming is a technique that performs source separation by means of spatial filtering using an array of sensors. Spatial filtering means that the beamformer discriminates between signals based on the physical location of the original sources. Beamformers can be adaptive or fixed, depending on whether the design of the beamformer relies on the signals observed by the array or not [10]. Beamforming techniques require knowledge of the sensor array configuration as well as the position of the sources. In contrast, BSS methods are able to recover the original sources from the observed mixed signals without any knowledge about sources, sensors or mixing process, but adding some constraints to the method [11, 9].

In this work, we focus on independent vector analysis (IVA) methods [12, 13, 14], included in the category of BSS techniques. IVA is an efficient multichannel source separation method and a variant of independent component analysis (ICA), which is a popular BSS method. However, IVA has at least two considerable drawbacks when we are dealing with speech and noise separations tasks; these drawbacks are related with the statistical source model assumed by IVA to achieve source separation. Firstly, the source model conventionally assumed in IVA does not provide an accurate representation for speech since it does not take into account its time-varying nature. Secondly, the conventional source model does not reflect the spectral

differences between the sources to be separated, which are speech and noise. The spectrum of speech signals is non-stationary, while background noise spectrum tends to be broadband and more stationary; but IVA assumes the same statistical model for all the sources. The present work is mainly motivated by these source model limitations of traditional IVA methods and our primary goal is to improve the source separation performance of IVA when separating speech from background noise. In this research, we hypothesize that we can achieve our goal by employing a new, improved source model for IVA that better represents speech and its differences to common noise signals. Specifically, in this work we assume a time-frequency-variant Gaussian distribution as IVA's source model. IVA implemented with the new proposed source model will be called *extended IVA* from now on.

A secondary goal of our research is to further improve the performance of extended IVA by introducing methods to reduce diffuse noise that may remain in the separated signals. This goal is motivated by the fact that removal of diffuse noise is a weakness of BSS methods in general and thus of IVA as well. With that in mind, we propose three post-processing filtering solutions to be concatenated with extended IVA. These solutions are in line with previous work on the subject [15] that have proved that this kind of approach enhances the performance of the multichannel source separation method under study. The goal of testing three different post-filtering solutions is to find a filter that combines the best with extended IVA.

Finally, in the present study we will address which evaluation measures on source separation and speech enhancement are more suitable in demonstrating the performance of the methods. Common measures for BSS usually neglect the perceptual aspects of the separated signals. In this work, we consider important to use measures that could represent not only the quantity of interference reduction, but also the perceptual quality of the separated signals. In this way, we can make more realistic conclusions about the methods under evaluation in real speech enhancement tasks.

The structure of this thesis is as follows. Theoretical background is first presented in Section 2; we start with a general view, by introducing the broader BSS topic and then more specifically the ICA methods. In Section 3, we introduce the central topic of the thesis, IVA methods, and our proposed approach to improve their source separation performance in speech and noise separation tasks, which is our main contribution in this work. In Section 4, the speech enhancement techniques used as support method for our approach are presented. Next, in Section 5, theory on post-processing of BSS methods for noise reduction is introduced, along with our contribution in relation to this. In Section 6, the experimental setup and evaluation methods of the source separation experiments performed in this work are presented. Results of these experiments are then shown in Section 7. The thesis concludes with Section 8, where the research done is summarized and some discussion points, such as possible future work of this research, are presented.

2 Independent component analysis

2.1 Blind source separation

BSS methods recover the source signals of interest from a set of observed mixed signals. The separation process is performed blindly, i.e. the source signals are unknown and so it is the mixing process. However, the so-called "blind methods" also include generic constraints, since otherwise the source separation problem would be unsolvable. There are several approaches to the BSS problem, depending on the constraint assumed. Some of the most important BSS approaches include ICA [16, 17], non-negative matrix factorization (NMF) [18, 19], and sparse component analysis (SCA) [20, 21]. In many BSS applications, the generic constraints are either complemented or replaced by specific constraints based on prior information such as the spectral content of the sources or the form of their probability densities. These specific constraints allow the design of more efficient and simpler source separation algorithms, which can be understood as semi-blind source separation (SBSS) instead of fully blind [11].

There are numerous applications for BSS such as speech separation, cross-talk removal in telecommunications, and analysis of brain imaging data like electroencephalographs (EEGs) or magnetoencephalographs (MEGs). Figure 1 shows an example of the BSS process for a speech separation task.

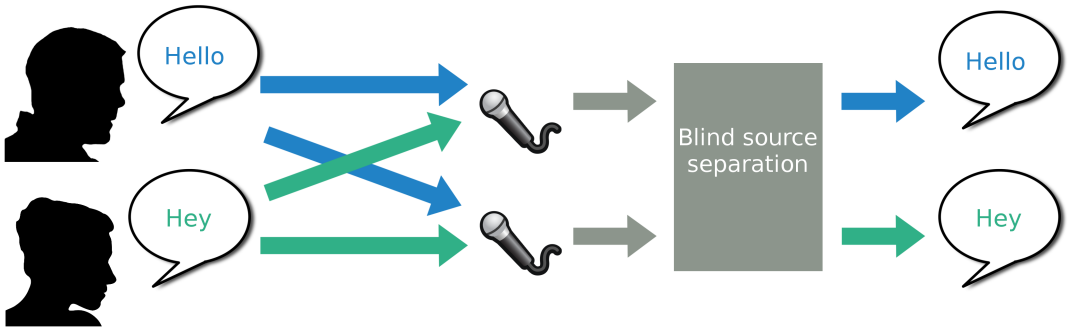

Figure 1: Blind source separation process.

2.2 Mixing models

The difficulty of a source separation task is highly dependent on the way the source signals are mixed in the physical environment in which they are transmitted [22]. Therefore, it is important that the mixing model assumed by the BSS method reflects as accurately as possible the real mixing process. The theory presented in this section is based on [23, 24, 25, 22], unless otherwise stated.

2.2.1 Instantaneous mixtures

Typically, the problem of BSS has been addressed with a simple mixing model called *instantaneous mixtures* and in consequence, most early BSS algorithms were designed according to this model. In instantaneous mixing, we assume that all signals arrive at the sensors at the same time. Then, each of the M observed signals $\{x_j(k)\}$, $1 \leq j \leq M$ consists of a weighted sum of N source signals $\{s_i(k)\}$, $1 \leq i \leq N$. In practice, the observed signals can be noisy, with the noise component corresponding to sensor noise or error noise from model inaccuracies [16][1]. In consequence, a noise term $v_j(k)$ is added to the mixing model. The instantaneous mixture model can be expressed then as

$$x_j(k) = \sum_{i=1}^{N} a_{ji} s_i(k) + v_j(k), \tag{1}$$

where $\{a_{ji}\}$ are the coefficients of the linear time-invariant (LTI) mixing system represented by the $M \times N$ mixing matrix A. The estimation of the sources is difficult when noise $v_j(k)$ is present. As a result, most of the research has neglected the noise term and applied a noise-free model [26]. In this work, we assume the simplified noise-free case for all the mixing models presented. Then, the instantaneous mixture model becomes

$$x_j(k) = \sum_{i=1}^{N} a_{ji} s_i(k), \tag{2}$$

which can be expressed in matrix notation as

$$\boldsymbol{x}(k) = A\boldsymbol{s}(k), \tag{3}$$

where $\boldsymbol{x}(k) = [x_1(k), \cdots, x_M(k)]^T$ and $\boldsymbol{s}(k) = [s_1(k), \cdots, s_N(k)]^T$; and ^{T} denotes vector transpose.

Before going further in the theory of mixing models, we will address how the number of source signals and sensors affect BSS. Under reasonable constraints, the BSS problem remains linear if the number of sensors is greater than or equal to the number of sources ($M \geq N$), and sources can be obtained by estimating mixing matrix A. In the determined case ($M = N$), the mixing matrix is square and the sources can be recovered by multiplying the observed signals with the inverse of the mixing matrix $W = A^{-1}$; W is commonly denoted as *demixing matrix*. In the overdetermined case ($M > N$), the separation task is still solvable with the pseudo-inverse.

Nevertheless, in situations in which the $M \geq N$ condition is not met, and the BSS problem becomes undetermined and thus cannot be solved linearly. In this case,

[1] The meaning of the noise term presented here differs with the definition of noise that we employ in the rest of the thesis: an acoustic signal from the environment that interferes with the signal of interest.

we are unable to obtain separated sources by simply inverting the mixing matrix. Now two different problems have to be solved, by first estimating the mixing matrix, and then the sources [16].

We focus here on the overdetermined and determined case ($M \geq N$). BSS is achieved by adjusting the coefficients w_{ij} of $N \times M$ demixing matrix W such that

$$y_i(k) = \sum_{j=1}^{M} w_{ij}(k) x_j(k), \tag{4}$$

is an estimate of the original source $s_i(k)$. In matrix notation, Equation (4) becomes

$$\boldsymbol{y}(k) = W\boldsymbol{x}(k), \tag{5}$$

where $\boldsymbol{y}(k) = [y_1(k), \cdots, y_N(k)]^T$. In Figure 2, the block diagram of the instantaneous mixture BSS process is shown. The simplicity of this model makes it useful for theoretical derivations, but it has limited applicability in real-world applications such as speech separation.

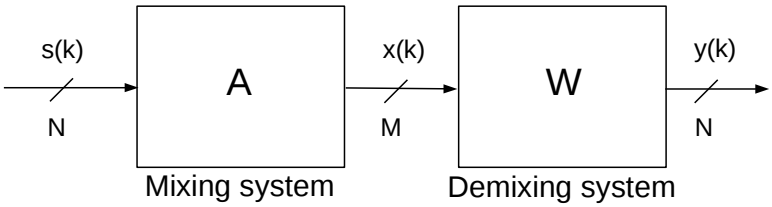

Figure 2: Block diagram of the instantaneous BSS task.

2.2.2 Convolutive mixtures

The instantaneous mixture model is rarely applicable for real-world situations, such as for acoustic mixtures, where the source signals are affected by the environment, and suffer from propagation delays, reverberation, etc. Therefore, it is often more appropriate to use models based on *convolutive mixtures*, in which the N source signals $\{s_i(k)\}$, $1 \leq i \leq N$, are mixed in a convolutive manner, since they are filtered by the impulse response of the environment through which they are propagated. Theoretically, the filters in the model should be of infinite length L, but in practice it is usually sufficient to assume finite length ($L < \infty$). Therefore, the signals $\{x_j(k)\}$ observed at the M sensors are given by

$$x_j(k) = \sum_{i=1}^{N} \sum_{l=0}^{L-1} a_{ji}(l) s_i(k-l), \tag{6}$$

where $\{a_{ji}(l)\}$ represent the impulse responses from source i to sensor j and are the coefficients of the $M \times N$ mixing matrices A_l. Therefore for convolutive mixtures, we have a discrete-time LTI mixing system $\boldsymbol{A}$, which is a matrix of linear filters[2] $\{A_l\}_{l=0}^{L-1}$ instead of scalars like it was for the case of instantaneous mixtures.

In convolutive separation, the separation or demixing system typically consists of a set of FIR filters $w_{ij}(l)$ of length L that are found blindly and produce N separated signals $\{y_i(k)\}$ as follows:

$$y_i(k) = \sum_{j=1}^{M} \sum_{l=0}^{L-1} w_{ij}(l) x_j(k-l), \tag{7}$$

where $w_{ij}(l)$ are the coefficients of the multichannel separation system $\boldsymbol{W}$, which is composed by $\{W_l\}_{l=0}^{L-1}$ with demixing matrices W_l of size $N \times M$. Figure 3 shows the BSS process for convolutive mixtures.

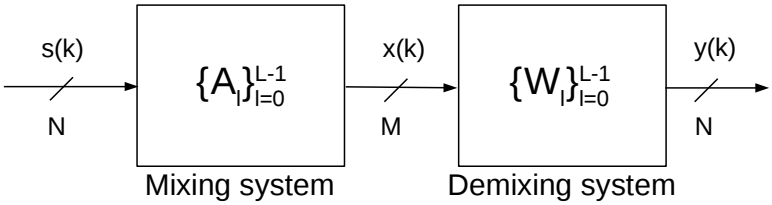

Figure 3: Block diagram of the convolutive BSS task.

In BSS, we would like, ideally, to separate the observed mixtures $x_j(k)$ and obtain the source signals $\{s_i(k)\}$, $1 \leq i \leq N$. However, this is a difficult task in case of signals "coloured" by the acoustical environment in which they propagate. A practical alternative goal is to obtain the convolved version of a source signal $s_i(k)$ observed at sensor j [24, 27]. In other words, the aim is to obtain the contribution of source $s_i(k)$ in channel j, called *source image* $s_{ji}^{img}(k)$, such that

$$x_j(k) = \sum_{i=1}^{N} s_{ji}^{img}(k) \tag{8}$$

For consistency, we will denote the estimate of $s_{ji}^{img}(k)$ as $y_{ji}^{img}(k)$.

2.3 Statistical constraints

ICA is one of the most important approaches in BSS. In ICA, we place certain restrictions and make assumptions, addressed here, in order to find the optimal

[2]Note that a matrix of filters, that is, a matrix with vector components, is marked in bold, in the same manner as vectors.

demixing matrix and in turn separate the components (what we designated before as sources). Firstly and most importantly, we assume the components are statistically independent. This contrasts with the mixed signals that are not independent since they have contributions from the same source signals. Therefore, to estimate the N independent components from the mixtures, our aim is to get statistically independent signals at the output. The components of vector $\boldsymbol{y} = [y_1, \cdots, y_N]^T$, considered random variables, are independent if the value of any of these components y_i is not affected by the occurrence of any of the others (y_j for $j \neq i$). Statistical independence can be defined mathematically in terms of probability density functions (p.d.f). The components of vector $\boldsymbol{y}$ are independent if and only if their joint probability $p_y(\boldsymbol{y})$ can be factorized as follows

$$p_y(\boldsymbol{y}) = p_{y_1}(y_1)p_{y_2}(y_2)\cdots p_{y_N}(y_N), \tag{9}$$

where $p_{y_i}(y_i)$ is the marginal p.d.f of component y_i [16].

Another assumption of ICA is that the components follow non-Gaussian distributions. Gaussian distributions can be considered "too simple" in the sense that their higher-order cumulants are zero-valued, and these statistics are fundamental for estimation of an ICA solution. Therefore, basic ICA is not applicable when the sources follow a Gaussian distribution[3] [16]. Nevertheless, in many cases we are mixing time signals, which have more structure than simple random variables. Additional statistics could be extracted from the time signal's structure in those cases and in turn, this information may enable the estimation of the components even in a situation in which basic ICA methods are unable; for example in case the sources are Gaussian distributed but correlated in time [16].

Finally, many times for simplicity the mixing matrix is assumed to be square, that is, we have as many sources as sensors. In consequence, we are assuming that the mixing matrix is invertible. This assumption, however, can be sometimes relaxed, we can deal with non-square cases (mentioned at the end of Section 2.2.1).

With the given assumptions, the separation problem is solvable, and we can find the demixing matrix up to some trivial ambiguities, discussed in Section 2.4.2 [16]. As we said earlier, to separate the components from the mixtures, we have to maximize the statistical independence of the estimated sources. For that, we first define a cost (or objective) function J, which is a measure of the statistical independence of all components in estimated source vector $\boldsymbol{y}$. Then, we minimize this cost function and find the optimal demixing matrix such that

$$\hat{W} = \arg\min_{W} J(W) \tag{10}$$

The variety of ICA algorithms can be differentiated mainly on the principle used for estimating the demixing matrix. Many of the most common methods are included in one of the following categories [16]:

[3]Strictly speaking, ICA model is still solvable if all sources are non-Gaussian except for one.

- **ICA by maximization of nongaussianity**. The algorithms in this category are motivated by the central limit theorem (CLT), which tell us, loosely speaking, that a sum of independent random variables has usually a distribution closer to Gaussian than any of the separate random variables. In this kind of algorithms, metrics such as kurtosis or negentropy are used to measure the nongaussianity of the estimated source signals. FastICA [28] is a popular ICA algorithm that belongs to this group.

- **ICA by maximum likelihood estimation**. ICA algorithms in this category employ the classic maximum likelihood (ML) estimation method to select the demixing matrix that gives the highest probability for the observed data. ML estimation requires prior knowledge of the densities of the source signals. Therefore, a common solution to this is to approximate the densities of the source signals by a family of (simple) parametric densities. In other words, we assume a statistical *source model* for the signals. Infomax is a well-known ICA algorithm that uses an estimation principle equivalent to the ML principle in case of source separation [29, 30].

- **ICA by minimization of mutual information**. This ICA approach, inspired by information theory, uses metrics like Kullback-Leibler (KL) divergence, which checks the difference between the real joint density of estimated source vector, $p_y(\boldsymbol{y})$, and the product of its approximated marginal densities (see Equation 9) to measure the independence of the estimated sources. Mutual information and maximum likelihood approaches are strongly connected and often lead to the same algorithms.

- **ICA by tensorial methods**. In this approach, higher-order cumulant tensors are used to estimate the sources. A popular ICA method included in this category is the JADE algorithm [31, 32].

2.4 Approaches in convolutive-mixture ICA

As discussed in Section 2.2, the difficulty of the source separation task is highly dependent on the mixing process of the source signals. In case of instantaneous mixtures, an instantaneous ICA algorithm can be employed directly to perform time-domain BSS. However, in case of convolutive mixtures, the task is more complex and ICA has to be extended to be applicable. Nevertheless, for convolutive mixtures, we can address the source separation problem by means of three different BSS approaches, depending on the domain where the operations are performed [33]. These three approaches are presented next.[4]

2.4.1 Time-domain approach

The first approach is time-domain BSS. Here the instantaneous ICA methods are directly extended to the convolutive case [33] and the demixing system, a matrix

[4]Without loss of generalization, we focus here on ICA instead of BSS.

of FIR filters, is computed in time domain (Equation 7) [34]. These algorithms give good separation results once the algorithm has reached convergence. This is because these methods evaluate quite accurately the statistical independence between the estimated source signals. However, convolutive-mixture ICA methods in time domain are more demanding than instantaneous ICA techniques, because of the need to compute convolution operations. This is especially true in case of reverberant mixtures that require the use of long filters to separate the mixed signals [33].

2.4.2 Frequency-domain approach

A second approach is frequency-domain BSS, where the time-domain signals are converted into frequency-domain using short-time Fourier transform (STFT) instead of global transform. The time-domain signals are split in blocks (typically called *windows* or *frames*) and a discrete Fourier transform (DFT) is computed for each block. Usually, smooth windowing functions are used, like a Hamming window, that tapers smoothly to zero at each end, and they overlap to some extent. Once the signal is transformed into frequency domain, the demixing system is estimated [34]. This is feasible since filtering the data before performing BSS does not change the mixing matrix, and applying Fourier transform does not modify the mixing matrix either [16].

In the frequency-domain approach, the convolutive mixtures in Equation (6) are approximated as one instantaneous mixture for each frequency bin [33]. In other words, we decompose the separation problem into several easier problems: now we have one complex-valued instantaneous ICA problem per frequency bin. Therefore, in the frequency-domain approach for convolutive mixtures, the dependencies between the source signals and observed mixtures are modeled as a linear mixing process. If we assume that N source signals are observed by M sensors, and that their STFT representations are obtained, the mixing process for frequency-domain BSS can be expressed as

$$\boldsymbol{X}_{\tau\omega} = A_{\omega}\boldsymbol{S}_{\tau\omega}, \tag{11}$$

where $\boldsymbol{X}_{\tau\omega} = [X_{1\tau\omega}, \ldots, X_{M\tau\omega}]^T$ denotes the Mx1 observation vector and $\boldsymbol{S}_{\tau\omega} = [S_{1\tau\omega}, \ldots, S_{N\tau\omega}]^T$ the Nx1 source vector at frequency channel ω in time frame τ, and A_{ω} is the unknown mixing matrix associated with channel ω. The vector component $X_{m\tau\omega}$ denotes the mixture observed with sensor m and $S_{m\tau\omega}$ the mth source signal at channel ω at time frame τ. The separated source signals $\boldsymbol{Y}_{\tau\omega}$ are obtained via the linear demixing process:

$$\boldsymbol{Y}_{\tau\omega} = W_{\omega}\boldsymbol{S}_{\tau\omega}, \tag{12}$$

where $\boldsymbol{Y}_{\tau\omega} = [Y_{1\tau\omega}, \ldots, Y_{N\tau\omega}]^T$ and W_{ω} is the demixing matrix at channel ω.

The main advantage of the frequency-domain approach for convolutive mixtures is that it reduces significantly the computational complexity of the task. Other

advantage is that any complex-valued instantaneous ICA algorithm can be directly applied to each frequency bin. Besides, the method can be computationally more efficient if we perform parallel computing of multiple frequency bins [33]. Finally, frequency-domain approaches have faster convergence [34].

Nevertheless, the frequency-domain approach also presents some drawbacks, two of which we address here. One problem is the circularity effect of the DFT representation of the convolutive mixtures. When the signals are converted to the frequency domain, the convolution becomes separate multiplications, one per frequency channel, but this is just an approximation that is only exact when source signal $s(k)$ is periodic, or equivalently, if the time convolution is circular. For linear convolution however we have errors at the frame boundaries of the STFT. A solution to this problem is to use the overlap-save method when going from time to frequency domain. However, a correct overlap-save algorithm is in some cases difficult to implement [23]. Parra and Spence [35] addressed also the circular/linear convolution problem and noted that a linear convolution can be approximated by a circular convolution, and therefore the errors due to the circular convolution can be reduced, if the frame length used in the STFTs is much larger than the length of the room impulse response (RIR). Specifically, the errors are reduced if the frame is at least two times longer than the RIR [36, 37]. However, the fixed resolution of STFTs implies that when long time frames are used, the number of samples in each frequency bin is small. As a result, the independence assumption of ICA collapses, since with poor frequency resolution it is difficult to get correct estimates of the statistics. Therefore, both short and long time frames fail at getting good results. An optimum frame size is determined then by the trade-off between having enough samples per frequency bin to get to estimate the statistics, and having a long enough time frame to cover the RIR [36].

Another problem of the frequency-domain approach is the permutation and scaling ambiguities that are inherent to the ICA solution. This means that even if we permute the rows of demixing matrix W_ω or multiply one row with a constant value, we will still have an ICA solution. This is expressed in mathematical form as

$$W_\omega \leftarrow \Lambda_\omega P_\omega W_\omega, \tag{13}$$

so updated W_ω is also an ICA solution for any permutation matrix P_ω and scaling (diagonal) matrix Λ_ω [33]. Permutation ambiguity involves that, if the order of the separated signals is not consistent across all the frequency channels, when the signals are transformed back to time domain, they will present contributions from different sources. This, in consequence, destroys the separation obtained in the frequency domain. On the other hand, the scaling ambiguity at each frequency bin causes an overall filtering of the estimated sources. Therefore, even when we would have perfect separation, the estimated sources would present different spectrum characteristics than the original sources, that is, distortion [23]. The scaling problem is easily solvable, but the permutation problem is more complex. Several solutions have been proposed to overcome the scaling problem, and a brief overview of them can be found in [23, 33]. One of the latest and successful approaches to solve

the permutation ambiguity was independent vector analysis (IVA), which will be explained in detail in Section 3.

2.4.3 Hybrid approach

Finally, the third approach to BSS is a hybrid of time-domain and frequency-domain approaches. The methods included here work in both time domain and frequency domain, taking advantage of the positive aspects of each domain. For example, some of these methods update the filter coefficients in frequency domain and evaluate the degree of independence between the sources in time domain. By evaluating the independence in time domain, we avoid de permutation problem. However, an important drawback of this approach is the increase of required computations of DFTs and inverse DFTs, due to the back and forth transformations of the signals between the time and frequency domains at each iteration. Therefore, the advantages of frequency-domain BSS are good enough to overlook its limitations, and it is in fact the most common approach in convolutive BSS [33].

3 Independent vector analysis

A multivariate variant of ICA, called IVA, was proposed some years ago to avoid the permutation ambiguity present in frequency-domain ICA [12]. Next, we present the connection of IVA to ICA and theoretical background of conventional IVA. Then, the IVA method proposed in this work, extended IVA, is explained in detail.

3.1 Conventional independent vector analysis

As we discussed in Section 2.3, to obtain estimates of the source signals in the ICA framework, we need to define a cost function J that measures statistical independence of the estimated sources, and we minimize it to find optimal demixing matrix W. Depending on the estimate principle used for separation, we have a different approach of ICA, and different forms of cost function. Mutual information (MI) could be seen as the canonical ICA cost function since it focuses on the key property of ICA, which is the independence of the sources [38]. The MI approach does not assume anything about the data other than the independence [16]; however, direct estimation of MI measure is computationally expensive in most cases. As a result, a common solution is to approximate MI by taking an approximation of the source signals densities and plug it into MI's definition in terms of entropies [16]. In other words, we are assuming a statistical source model for each source signal.

IVA methods originated as a solution to the permutation ambiguity problem of frequency-domain ICA [12]. The key difference between ICA and IVA is the source model: IVA employs multivariate source models, while ICA uses univariate models instead. In IVA, all the frequency components are modelled as stochastic vector variables (a source-wise vector $\tilde{\boldsymbol{Y}}_{m\tau} = [Y_{m\tau 1}, \ldots, Y_{m\tau\Omega}]^T$, where Ω is the total number of spectral channels and m is the number of source channel) and the sources are separated vector-wise instead of frequency wise, as it was the case for ICA. Therefore, the dependencies between the frequency channels are represented via the source model, a multivariate probability density function $p_y(\tilde{\boldsymbol{Y}}_{m\tau})$, and the method estimates the source signals by looking for statistically independent sources while keeping the dependencies between the spectral channels. With this, the permutation ambiguity is avoided.

From what we have explained, it is clear that in the IVA framework a cost function for multivariate random variables is needed. The cost function of IVA to be minimized is, in terms of the MI approach, [13, 14] given by

$$J_1 = \sum_m H(\tilde{\boldsymbol{Y}}_m) - H(\tilde{\boldsymbol{Y}}), \tag{14}$$

where $H(\tilde{\boldsymbol{Y}}_m)$ is the differential entropy of $\tilde{\boldsymbol{Y}}_{m\tau}$, the source-wise vector of the mth estimated source signal, and $H(\tilde{\boldsymbol{Y}})$ is the joint entropy of $\tilde{\boldsymbol{Y}}_\tau = [\tilde{\boldsymbol{Y}}_{1\tau}, \ldots, \tilde{\boldsymbol{Y}}_{N\tau}]$. J_1 is the expression of mutual information, extended to measure dependency between multivariate random variables . Also, J_1 is equivalent to the KL divergence of the real joint probability of estimated source signals $p_y(\tilde{\boldsymbol{Y}}_\tau)$ and the factorized density

q $= p_y(\tilde{\boldsymbol{Y}}_{1\tau})p_y(\tilde{\boldsymbol{Y}}_{2\tau})\dots p_y(\tilde{\boldsymbol{Y}}_{N\tau})$, where $p_y(\tilde{\boldsymbol{Y}}_{m\tau})$ is the marginal density of mth estimated source.

The entropy of a linear transformation $\boldsymbol{r} = D\boldsymbol{b}$ is given by $H(\boldsymbol{r}) = \log\det|D| + H(\boldsymbol{b})$, where $\boldsymbol{r}$ and $\boldsymbol{b}$ are vector random variables and D is a matrix. If we express the estimated source signals in function of the observed signals (Equation 12) and use the aforementioned property, cost function J_1 becomes

$$J_1 = \sum_m H(\tilde{\boldsymbol{Y}}_m) - \sum_\omega \log\det|W_\omega| - H(\tilde{\boldsymbol{X}}), \tag{15}$$

where $H(\tilde{\boldsymbol{X}})$ is the joint entropy of $\tilde{\boldsymbol{X}}_\tau = [\tilde{\boldsymbol{X}}_{1\tau}, \dots, \tilde{\boldsymbol{X}}_{M\tau}]$, with (source-wise) observed vectors $\tilde{\boldsymbol{X}}_{m\tau} = [X_{m\tau 1}, \dots, X_{m\tau\Omega}]^T$. $H(\tilde{\boldsymbol{X}})$ is a constant, so we can discard it to simplify the cost function expression. Finally, if we express the entropy as an expectation, we get one of the most common forms of IVA cost function:

$$J_1 = \sum_m \frac{1}{T}\sum_\tau G(\tilde{\boldsymbol{Y}}_{m\tau}) - \sum_\omega \log\det|W_\omega|, \tag{16}$$

where $G(\tilde{\boldsymbol{Y}}_{m\tau})$ is called *contrast function*, computed as $G(\tilde{\boldsymbol{Y}}_{m\tau}) = -\log p_y(\tilde{\boldsymbol{Y}}_{m\tau})$, and T is the total number of frames. The demixing matrices are iteratively estimated by minimizing this objective function with regards to W_ω. Minimizing Equation (16) is equivalent to ML estimation.

As discussed above, the source models in IVA take into account the dependency between the spectral channels for each source. The dependency is modeled by assuming a multivariate probability density function $p_y(\tilde{\boldsymbol{Y}}_{m\tau})$ as source model, for a source-wise vector $\tilde{\boldsymbol{Y}}_{m\tau}$. The conventional source models in IVA are spherical, time-invariant, and super Gaussian distributions [12, 13], such as

$$p_y(\tilde{\boldsymbol{Y}}_{m\tau}) \propto \exp\left\{-K\sqrt{\|\tilde{\boldsymbol{Y}}_{m\tau}\|_2^2}\right\}, \tag{17}$$

where K is a time-invariant constant and $\|\cdot\|_2$ denotes the L_2 norm of a vector.

3.2 Extended independent vector analysis

In this work, we focus on IVA in speech and noise separation tasks, and propose solutions to overcome two important limitations of IVA in these kind of tasks. Firstly, the models commonly used in IVA are time-invariant and therefore they do not model the time-varying nature of speech. In addition, baseline IVA typically assumes the same source model for all the sources. In case of speech and noise separation tasks, this assumption is not correct, since the spectra of these sources have very different characteristics. The spectrum of a speech signal is non-stationary and it is characterized by its pitch and formant frequencies. In contrast, background noise usually has a broad band spectrum and might be temporally stationary.

In the present work, we propose an extension of IVA with a new source model more suitable for speech and noise separation. From now on, we call the proposed

method as extended IVA. The new source model takes into account the issues we have described above, and that are neglected by conventional source models. Therefore, this source model no longer follows a spherical, time-invariant, and super Gaussian distribution, like the one shown in Equation (17). In contrast, the source model assumed now is a time-frequency variant Gaussian distribution, such as

$$p_y(Y_{m\tau\omega}) \propto \frac{1}{\sigma^2_{m\tau\omega}} \exp\left\{-\frac{Y^2_{m\tau\omega}}{\sigma^2_{m\tau\omega}}\right\}, \tag{18}$$

where $\sigma^2_{m\tau\omega}$ is the variance of mth source at time frame τ and frequency ω.

The model proposed here includes the temporal power variations of the sources. IVA has been evaluated before with time-variant source models in [39, 40], where distribution variances were assumed constant across frequency channels. However, in our case, the distribution variances $\sigma^2_{m\tau\omega}$ have a different value for each frequency channel. Besides, we assume to have available a single-channel source separation method. Then, the variances $\sigma^2_{m\tau\omega}$ are computed as

$$\sigma^2_{m\tau\omega} = |\hat{S}_{m\tau\omega}|^2, \tag{19}$$

where $\hat{S}_{m\tau\omega}$ is the output from the single-channel source separation method for the mth source at time frame τ and frequency ω. Most single-channel source separation methods rely on the spectral differences between the sources. Therefore, by plugging information from a single-channel separation method into our source model, extended IVA takes also into account the differences between sources. With the proposed source model, the source separation performance of IVA is expected to improve in speech and noise separation tasks.

Using the proposed source model, the cost function of conventional IVA (Equation 16) is transformed into the following expression,

$$J_2 = \sum_{\omega}\left(\sum_{m}\frac{1}{T}\sum_{\tau}\frac{\|\boldsymbol{w}^H_{m\omega}\boldsymbol{X}_{\tau\omega}\|^2_2}{\sigma^2_{m\tau\omega}} - \log\det|\,W_\omega|\right), \tag{20}$$

where $\boldsymbol{w}^H_{m\omega}$ is the mth row of the demixing matrix W_ω and ^{H} denotes Hermitian transpose. J_2 is the cost function for extended IVA.

3.3 AuxIVA: IVA based on an auxiliary function technique

The extension of IVA evaluated on this work was implemented on AuxIVA, an IVA method based on an auxiliary function technique [41]. Typically, IVA algorithms, which compute the demixing matrix, are based on natural gradient updates [12, 13, 14]. This type of algorithms have a trade-off between the convergence speed and stability. The approach in AuxIVA, first developed in the ICA framework [42] and later extended to IVA, presents more effective update rules [41]. AuxIVA method involves two alternative update steps. The update rules in case of extended AuxIVA

are as follows. First the weighted covariances matrices $V_{m\omega}$ are once calculated for all ω as

$$V_{m\omega} = \frac{1}{T} \sum_{\tau} \left(\frac{\boldsymbol{X}_{\tau\omega} \boldsymbol{X}_{\tau\omega}^{H}}{\sigma_{m\tau\omega}^{2}} \right) \tag{21}$$

Then the demixing matrices are updated. No close form for updating simultaneously $w_{m\omega}$ in Equation (20) has been proposed yet. Therefore, we consider an update of only $w_{m\omega}$ while keeping the other $w_{l\omega}(l \neq m)$ fixed. Then, the demixing matrix update rules, for all ω, are

$$\boldsymbol{w}_{m\omega} \leftarrow (W_{\omega} V_{m\omega})^{-1} \boldsymbol{e}_{m}, \tag{22}$$

$$\boldsymbol{w}_{m\omega} \leftarrow \frac{\boldsymbol{w}_{m\omega}}{\sqrt{\boldsymbol{w}_{m\omega}^{H} V_{m\omega} \boldsymbol{w}_{m\omega}}}, \tag{23}$$

where $\boldsymbol{e}_{m}$ is a unit vector with the mth element unity $\boldsymbol{e}_{m} = [0, \ldots, 1, \ldots, 0]$. The update rules are applied iteratively until convergence is achieved.

A variant of AuxIVA for stereo signals exists; it achieves faster convergence than the general AuxIVA method [43]. In this work, we used general AuxIVA for the empirical evaluations.

4 Speech enhancement techniques as support methods

Next, we introduce briefly the speech enhancement techniques used in the present work as support methods to obtain source estimates for the source model of extended IVA.

4.1 REPET SIM

REpeating Pattern Extraction Technique (REPET) is a speech enhancement technique that separates repeating background from non-repeating foreground in a mixture. The separation is based on finding the repeating patterns in an audio mixture, deriving the underlying repeating models and finally extracting the repeating background by comparing the models to the mixture. REPET SIM is a generalization of the REPET method that can handle non-periodically repeating structures. In this case, similarity matrices are used to identify the repeating elements in the mixture, based on the assumption that the background, noise for example, is dense and low-ranked, while the foreground, speech for example, is sparse and varied [44]. According to what we explained in Section 3.2, source estimates to use in Equation (19) are obtained with a single-channel source separation method, since most of these methods rely on the spectral differences between the sources. REPET SIM is a multichannel source separation technique, but it can be applied also to single-channel data. Since this technique relies on differences between the frequency spectrums of the sources, as single-channel source separation techniques do, REPET SIM is a good option to use for computation of the source estimates.

4.2 Spectral subtraction

Spectral subtraction is a single-channel speech enhancement method used for noise reduction. Like many single-channel methods, spectral subtraction relies on the spectral differences between the sources, and therefore it is also a good choice to obtain the source estimates. In these methods, the enhancement of the speech signal corrupted by noise is achieved by subtracting an estimate of the noise spectrum from the noisy speech spectrum. In the present work, the spectral subtraction implementation used is based on [45]. This approach differs slightly from the basic principle of spectral subtraction in two main ideas. First, now an overestimate of the noise spectrum is subtracted from the noisy speech spectrum; this means that we subtract a factor (α) times the noise spectrum estimate, where α is a number larger than one. The value of α varies from frame to frame, according to the signal to noise ratio (SNR) measured; the larger the SNR value, the smaller the value of α is. Second, the resultant spectrum is lower-bounded by a minimum value called *spectral floor*. With the aforementioned changes, the approach eliminates "musical noise" that the original spectral subtraction method introduces in the enhanced signal [45].

5 Post-processing of multichannel source separation

5.1 Background

BSS methods have a limited capability to reduce diffuse noise. In a diffuse noise field, the noise propagates from many directions and, because of the acoustic properties of the environment, it is perceived at the sensors as coming from all directions. Diffuse noise field has proven to be a reasonable model for many real-life noise environments, like for example babble noise in cafeterias and car noise environments [46, 47].

The diffuse noise reduction limitation also occurs in other multichannel source separation methods. For example, adaptive beamformers (ABF), such as minimum variance distortionless response (MVDR) beamformers, have also a limited noise reduction capacity when the noise field is diffuse [48, 15]. Frequency-domain BSS has been proved to be equivalent to frequency-domain ABFs and as ABF, BSS mainly removes the sound coming from the direction of interference. It must be noted, though, that ABF does not involve an assumption of independency of the source signals as BSS does. Therefore, the source separation performance of ABF is not affected in case the independency assumption collapses. This means that the performance of BSS is upper bounded by that of ABF [49].

The research on noise reduction has attracted much interest in the past years, and part of it has been focused on multichannel speech enhancement methods with post-filtering [15]. Post-filtering methods can be divided into two main groups: single-channel post-filters and multichannel post-filters. Multichannel Wiener filter (MWF) is the best possible linear filter for multichannel noise reduction of broadband inputs in the minimum mean squared error (MMSE) sense [15]. Simmer et al [15] proved that, assuming the target signal and noise are mutually uncorrelated, MWF can be factorized into a MVDR beamformer followed by a single-channel Wiener post-filter. A MVDR beamformer coupled to a single-channel Wiener post-filter produces a higher output SNR than the MVDR beamformer alone.

5.2 Post-processing in this work

In the present work, we focus on single-channel post-filtering as an approach to further improve the source separation performance of IVA when diffuse noise is present. Our approach originates from the theoretical principle of MWF and its factorization into two stages [15] (see Section 5.1). We present an analogous setup, where the proposed extended IVA method is concatenated with a time-variant single-channel post-filter (Figure 4). In this setup, the source estimates $Y'_{m\tau\omega}$ are calculated based on the multichannel estimates $Y_{m\tau\omega}$ from extended IVA (with source estimates $\hat{S}_{m\tau\omega}$ information from single-channel source separation) as

$$Y'_{m\tau\omega} = H_{m\tau\omega} Y_{m\tau\omega} \tag{24}$$

where $H_{m\tau\omega}$ is the STFT representation of the single-channel post-filter applied on the mth source estimate in time frame τ and frequency channel ω. In this work, we evaluate three time-variant post-filters $H_{m\tau\omega}$ based on the multichannel source

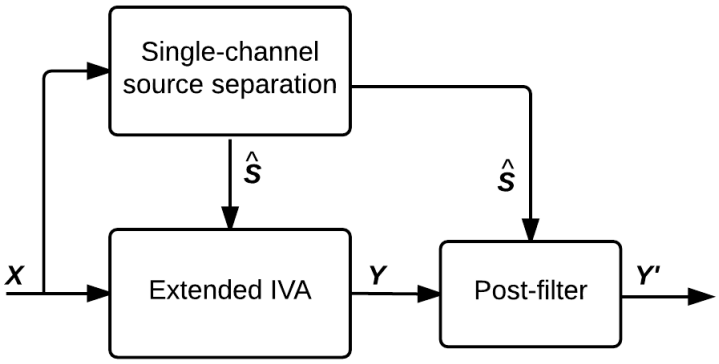

Figure 4: Block diagram of the setup proposed: the multichannel source separation system proposed earlier, extended IVA with single-channel source separation, concatenated with one single-channel post-filter per source. $\boldsymbol{X}$ denote the observation vector and $\boldsymbol{Y'}$ the estimated sources vector.

estimates $Y_{m\tau\omega}$ and source estimates $\hat{S}_{m\tau\omega}$ calculated with a single-channel source separation method. For reference, we also evaluated a setup of extended IVA plus single-channel source separation. Next, we present the three single-channel post-filter approaches we propose.

Wiener post-filters

The first two post-filters proposed are both Wiener filters. We chose this type of filter based on the MWF factorization proved in [15] (see Section 5.1). A Wiener filter is designed such that it meets the MMSE criteria, and according to the Wiener-Hopf equation, the general expression in the frequency domain is ϕ_{ys}/ϕ_{yy}, where ϕ_{ys} is the cross power spectrum density (CSD) of observed signal $y(k)$ and target signal $s(k)$ and ϕ_{yy} is the power spectral density (PSD) of $y(k)$. When the target signal and noise are uncorrelated, the CSD term can be reduced to $\phi_{ys} = \phi_{ss}$ and the PSD term in the denominator can be expressed as $\phi_{yy} = \phi_{ss} + \phi_{nn}$, where ϕ_{ss} is the PSD of the desired signal and ϕ_{nn} the noise PSD [15, 50, 51]. Given these simplifications of the general expression, the first post-filter is calculated as

$$H^{(1)}_{m\tau\omega} = \frac{|\hat{S}_{m\tau\omega}|^2}{|\hat{S}_{m\tau\omega}|^2 + |N_{m\tau\omega}|^2}, \tag{25}$$

where the noise $N_{m\tau\omega}$ is calculated as $N_{m\tau\omega} = Y_{m\tau\omega} - \hat{S}_{m\tau\omega}$.

In the second post-filter, we use the simplified form of the general expression's numerator, obtained from the assumption of uncorrelated signal and noise. However, we keep the denominator as it is given by the general expression. In consequence, for guaranteeing that the post-filter amplitudes fall between the range [0, 1], simply

clipping is applied. Then, this filter can be represented as

$$H^{(2)}_{m\tau\omega} = \min\left\{\frac{|\hat{S}_{m\tau\omega}|^2}{|Y_{m\tau\omega}|^2}, 1\right\}. \tag{26}$$

Amplitude replacing post-filter

Most single-channel source separation methods estimate the sources with time-frequency masking that modifies the amplitudes while keeping the phase of the observed mixed signals. It has been proven that the phase is as important as the amplitude for correct separation of the signals [52], and using simply the phase of observed signals as the phase of estimated source signals causes problems in that traces of interfering signals remain in the source signals. In contrast, extended IVA does not have these problems as it estimates the phase as well as the amplitude of the separated signals. With this in mind, our third post-filter approach combines the amplitude estimation from single-channel source separation and the phase estimation from extended IVA. That is, when the post-filter $H^{(3)}_{m\tau\omega}$, given by

$$H^{(3)}_{m\tau\omega} = \frac{|\hat{S}_{m\tau\omega}|}{|Y_{m\tau\omega}|}, \tag{27}$$

is applied to the multichannel estimates $Y_{m\tau\omega}$ from extended IVA (Equation 24), the amplitude of estimates $Y_{m\tau\omega}$ are substituted by the amplitudes of the source estimates $\hat{S}_{m\tau\omega}$ from single-channel separation, while keeping the phase from extended IVA.

6 Experimental work

In this section, the experiments performed in the current work are presented. The experiments focus on BSS utilized for speech separation and speech and noise separation. The speech and noise separation task involves diffuse noise (see Section 6.1.2), which is challenging for BSS methods, as discussed in Section 5.1. For that reason, prior to the speech and noise separation experiments, we performed a speech separation experiment with localized sources, the optimal sources IVA can handle, to first test that our proposed extended IVA method was working. In other words, with this experiment we would be able to validate our hypothesis that a new improved speech source model for IVA would lead to an improvement of its source separation performance on speech data. The new source model is designated *improved* since this model would represent speech more accurately than the original source model does.

The section is organized as follows. First, the data employed in the experiments is described in Section 6.1. Then, the setup of the methods applied in our work is presented on Section 6.2. Finally, the evaluation metrics used are shown in Section 6.3.

6.1 Data

6.1.1 Speech separation experiment

We first evaluated extended IVA on a speech separation task with data from the Signal Separation Evaluation Campaign (SISEC) 2008 [53]. The samples employed were a selection from the development data set of the undetermined speech and music mixtures task. This set contains mixtures of female speech, male speech and music. The mixtures originate from three or four sources that are observed at two microphones (stereo microphone). Since it is a development set, this set also includes the source signals and source images corresponding to the mixtures. A source image is the contribution of the corresponding source signal to the mixtures observed at the sensors (more detailed explanation at the end of Section 2.2.2). The data set includes different mixing conditions: three kinds of mixtures (live recording, its artificial counterpart, synthetic convolution, and instantaneous mixture), two reverberation times (130 ms and 250 ms), and two cases for microphone spacing (5 cm and 1 m). The duration of all the samples in the set is 10 s. and their sampling frequency is 16000 Hz.

For the experiment, all the speech samples, female and male speech, corresponding to live recording mixtures that originate from four sources were selected. However, the extended IVA method proposed in this work is implemented on the AuxIVA algorithm [41], which works under the assumption of determined or overdetermined case, that is, the number of microphones is equal to or larger than the number of sources. For this reason, in the experiment the number of sources and microphones were both set to two, and we created our own test samples instead of using the original SISEC samples. The mixtures (or test samples) were created by mixing the source images observed at each microphone such that each mixture consisted of 2

source signals instead of 4 sources in the original SISEC data. All possible source signals' permutations were obtained for each mixture condition (reverberation time and microphone spacing); with a total of 48 mixtures for evaluation.

6.1.2 Speech and noise separation experiments

Extended IVA was also evaluated on a block of experiments focused on speech and noise separation, with part of the material from SISEC 2013 [54]. Specifically, we used the development data set of the two-channel mixtures of speech and real-world background noise task. In this case, we employed all the development set samples for evaluation. This set consists of nine stereo recordings of a speech source that is contaminated by real-world diffuse noise. The diffuse noise was recorded in three kinds of public environments: a subway car, cafeterias and squares. Apart from the nine stereo mixtures, the set also includes the corresponding source signals (speech and noise) and source images. All samples of this data set have a duration of 10 s. and their sampling frequency is 16000 Hz. Also in these experiments, extended IVA assumes 2 sources and microphones, but now the data set employed attains this. Therefore, we did not need to regenerate data as in the previous experiment, and used for evaluation in these experiments the original 9 mixtures of the development set.

6.2 Experimental setup

6.2.1 IVA settings

The extended IVA method evaluated in this work was implemented on AuxIVA [41]. We addressed in Section 2.4.2 the STFT frame length effects in the source separation performance. Too short or too long STFT time frames may fail at getting good source separation results, since even though each case involves some benefits they also bring some problems, which will vary depending on the conditions of the BSS task at hand. Therefore, we conducted experiments over a range of STFT frame lengths in order to find an value: IVA was applied on STFTs computed on 512, 1024, 2048, and 8192-sample Hamming windows with 50% overlap. An identity matrix was employed as initial value for the demixing matrix, and the algorithm was iterated 20 times to ensure convergence. In addition, the source variances used in extended IVA were computed based on source estimates as indicated in Equation (19). For the speech separation task, we evaluated extended IVA with variances calculated based on oracle information, with $\hat{S}_{m\tau\omega} = S_{m\tau\omega}$. The oracle variances were computed from the true source images. This is because the IVA method used in this work, AuxIVA, gives at its output the estimate of the source images observed at the specified microphone. By using the oracle variances, extended IVA is not dependent on the performance of the support method used to obtain the source estimates needed to compute the variances. In consequence, as stated already in Section 6, this experiment is used to validate the improved source model hypothesis (see Section 1). In addition, we can determine the upper performance limit for extended IVA in the given conditions.

In case of the speech and noise separation experiments, the speech and noise source estimates were calculated with two different speech enhancement methods, REPET SIM (Section 4.1) and spectral subtraction (Section 4.2). We used two different methods to ensure the final source separation results would not be directly dependant on the method used for obtaining the source estimates. In addition, we also ran the speech and noise separation experiments using oracle variances for extended IVA, as we did for the speech separation task. This was done to have an idea of the potential that extended IVA has in more challenging situations like speech in diffuse noise - the condition in these experiments. Finally, we note that while in these experiments we get both speech and noise as outputs of IVA, we only evaluate speech, which is the focus of this work.

6.2.2 REPET SIM settings

REPET SIM is one of the speech enhancement methods whose source estimates were used to compute the source variances (see Section 4.1 for more details on the method). In our experiments, the parameters of the REPET SIM algorithm were fixed as follows: minimum similarity threshold between a repeating frame and the given frame $t = 0$, minimum distance between two consecutive repeating frames $d = 0.1$ seconds, maximum number of repeating frames $k = 20$, and maximal past and future buffer size $buf = [2, 2]$ seconds. The rest of parameters were set to default values. All parameters' values were based on the optimal parameters proposed in [44] for the data under evaluation (detailed in Section 6.1.2). In the algorithm script, the similarity parameters were taken as input argument in the form $par = [t, d, k]$.

6.2.3 Spectral subtraction settings

In our experiments, we also employed the spectral subtraction method implemented in Matlab toolbox VOICEBOX [55] to obtain the source estimates used for computing the source model variances. In the experiments, we employed the default parameters of the spectral subtraction script *specsub*, which was modified to output the noise signal in addition to the speech signal that was the only output in the original script. The noise estimate was computed by applying to the spectrum of the mixture a soft-decision mask that is complementary to the one *specsub* computes to obtain the speech estimate.[5]

6.3 Evaluation

6.3.1 Standard measures for BSS performance evaluation

About ten years ago not much attention was paid to evaluation metrics for BSS in speech applications. There was research on the subject, such as [56], but most literature was focused on the development of new BSS algorithms. Very different evaluation metrics were used on the research papers published at the time due to

[5] A soft-decision mask has elements ranging $[0, 1]$; by complementary mask we mean that if we have soft-decision mask M_1, the complementary soft-decision mask will be $M_2 = 1 - M_1$.

the lack of a "benchmark", not only for BSS in speech-related tasks, but also for BSS algorithms in general. In consequence, the comparison between algorithms was difficult [57].

Later, efforts have been made to establish a standardized evaluation framework for BSS, such as the work from Vincent et al. [58]. In this work, a new performance criterion was proposed for evaluation of BSS in audio signals. This criterion consisted of four energy ratios, in which the estimated source signal y_i is compared to the corresponding true source s_i. These measures do not take into account the permutation indeterminacy of BSS. Therefore, if necessary, y_i would have to be compared with all the sources $\{s_i\}$, $1 \leq i \leq N$. Besides, the criterion involves two assumptions. First, the true source signals and noise signals (if any) are known. Second, the user chooses a family of allowed distortions $\mathcal{F}$ for the target signal according to the application, but independently of the mixture or algorithm used. This means that we allow our target signal $s^{target}(k) = f(s(k))$ to be a version of source signal $s(k)$ modified by an allowed distortion $f \in \mathcal{F}$. An advantage of this criterion is that the mixing system and demixing technique do not need to be known.

The aforementioned criterion was developed for evaluation of estimated single-channel source signals. Inspired on it, a similar criterion was developed that evaluates estimated source images [59]. Both criteria are computed on two steps. First, decomposition of the estimated signal. In case of the source image criterion, the source image from source i observed at microphone j $y_{ji}^{img}(k)$, is decomposed as

$$y_{ji}^{img}(k) = s_{ji}^{img}(k) + e_{ji}^{spat}(k) + e_{ji}^{interf}(k) + e_{ji}^{artif}(k), \tag{28}$$

where $s_{ji}^{img}(k)$ is the true source image and $e_{ji}^{spat}(k)$, $e_{ji}^{interf}(k)$ and $e_{ji}^{artif}(k)$ are the error components representing the spatial (or filtering) distortion, interference and artifacts, respectively. In the second step, the energy ratio measures are computed. The energy ratio measures proposed in the source image criterion, expressed in dB, evaluate spatial distortion, interference and artifacts. They are the source image to spatial distortion ratio (ISR_i), the signal to interferece ratio (SIR_i) and the signal to artifacts ratio (SAR_i), respectively. Also, the signal to distortion ratio (SDR_i) measure encompasses all the previous error terms: spatial distortion, interference and artifacts. All these measures are similar to the ones from the previous criterion [58], but now the target signal is split in two terms, $s_{ji}^{img}(k)$ and $e_{ji}^{spat}(k)$.

The source image criterion has the advantage that it allows the evaluation of source signals that cannot be represented as single-channel signals, such as common audio signals that are typically presented in stereo (two-channel) format: radio, television, music CDs and MP3s, etc. Besides, potential gain or filtering indeterminacies about the estimated single-channel source signal $s_i(k)$ disappear when the source image is considered for evaluation instead [27].

For convolutive mixtures, source to interference ratio (SIR) is a common metric to report the performance of BSS algorithms [23], as it is the source to distortion ratio (SDR). In the present work, we used SDR since it is a summarizing metric in contrast with the rest of energy ratios that deal with specific error components. Specifically, the source image counterpart of SDR, which is SDR_i, was used for

evaluation since as we mentioned already in Section 6.2.1, AuxIVA gives estimates of source images. Another reason to use this measure was that SDR_i was one of the evaluation metrics of SISEC 2008 and SISEC 2013, campaigns from which the two data sets employed in our experiments are from. The SDR_i metric is defined as

$$SDR_i = 10 \log_{10} \frac{\sum_{j=1}^{M} \sum_k s_{ji}^{img}(k)^2}{\sum_{j=1}^{M} \sum_k (e_{ji}^{spat}(k) + e_{ji}^{interf}(k) + e_{ji}^{artif}(k))^2} \tag{29}$$

In this work, SDR_i measures were computed with the BSS Eval Matlab toolbox [59, 60].

Even though the evaluation criteria based on energy ratios are commonly used, they present the following two limitations. First, the numerical precision of the measures is lower for high-performance values than for low ones. For example, in case we have a high value of SDR, this means that the denominator of this energy ratio is very small. In consequence, small amplitude errors in the numerator, such as signal quantization errors, will cause large SDR deviations. Second, these measures cannot explain certain properties of the auditory properties. For example, the energy ratios, at high values, have limited auditory significance; and the ratio SDR does not measure the total perceived distortion [58].

6.3.2 Perceptual measures

Quality evaluation of the separated signals should take into account the final task where the signals are to be used. In applications where the final result is going to be listened by humans, perceptual quality is much more important than perfect reconstruction of the original waveform. Speech quality assessment using listening tests with human subjects is often accurate and reliable when it is performed under stringent conditions, however it involves high costs in time and resources [61]. Consequently, several objective quality measures have been proposed to predict the subjective quality of speech [62]. Most of these measures were developed originally in the telecommunication field for evaluation of the distortion introduced by speech codecs and/or communication channels. Therefore, at first, the suitability of these measures on predicting subjective quality of enhanced speech was not clear. A recent study [61] has tested the correlation of several of these objective measures with subjective listening tests, when evaluating enhanced speech. Of the seven objective measures under study, the perceptual evaluation of speech quality (PESQ) measure [63] presented the highest correlation to the subjective assessments. However, a drawback of PESQ is its computational complexity. The same study [61] reported that the objective measure called frequency-weighted segmental SNR (fwSNRseg) performed nearly as well as PESQ at much less computational cost. In the present work, we decided to employ fwSNRseg as a complementary measure to SDR_i, to take into account the perceptual aspects of the separated signals in the evaluations.

fwSNRseg is computed as follows

$$fwSNRseg = \frac{10}{T}\sum_{\tau=1}^{K}\frac{\sum_{\omega=1}^{\Omega} wgt(\tau,\omega)\log_{10}\frac{|S_{\tau\omega}|^2}{|S_{\tau\omega}-\widehat{S}_{\tau\omega}|^2}}{\sum_{\omega=1}^{\Omega} wgt(\tau,\omega)}, \tag{30}$$

where $wgt(\tau,\omega)$ is the weight in the mel frequency band ω at time frame τ, T is the total number of time frames, Ω is the total number of mel frequency channels, $S_{\tau\omega}$ and $\hat{S}_{\tau\omega}$ are the clean signal spectrum and the enhanced signal spectrum, respectively, at time frame τ and mel filter channel ω. Weighting function $wgt(\tau,\omega)$ is computed as

$$wgt(\tau,\omega) = |S(\tau,\omega)|^{\gamma}, \tag{31}$$

where γ is a power exponent that can be varied for maximum correlation. In the present work, we used a exponent value of $\gamma = 0.2$, as proposed in [61]. The spectra of clean and enhanced signals were obtained by first computing their STFTs with Hamming windows of 25 ms, shifts of 5 ms between adjacent frames, and 1024-point Fourier transforms. The signal bandwidth was then grouped into 21 bands using a mel filter bank, which was computed using VOICEBOX Matlab toolbox [55]. Besides, for computing the average on Equation (30), the SNRs obtained at each time frame were limited to the range of -10 to 35 dB as in [61]. We chose the above settings based on [64], where they proved to work for perceptual evaluation of enhanced signals.

7 Results

7.1 Speech separation

First, the results of the speech separation experiments conducted on extended IVA are presented in Tables 1 and 2. In this experiment we computed extended IVA with oracle variances and compared with baseline IVA. For both of the evaluation metrics employed, SDR_i and fwSNRseg, the average over all sources and trials was computed and used as final evaluation value. The task was evaluated with four different STFT frame lengths in IVA: 512, 1024, 2048 and 8192 samples. Extended IVA with oracle variances shows an improvement over baseline IVA for all frame length cases. When comparing baseline IVA results against extended IVA results, we also note that the optimal frame length for baseline IVA is 2048 samples, while for extended IVA it is 8192 samples. Besides, the results suggest that extended IVA is more sensitive to frame length, since the variation of results between 512 and 8192 frame lengths is larger for extended IVA than for baseline IVA.

Table 1: SDR_i results [dB] for baseline IVA and extended IVA (oracle) in four STFT frame length [samples] cases. The best SDR_i value in each case is underlined.

	512	1024	2048	8192
Baseline IVA	3.30	4.20	4.84	3.48
Extended IVA (oracle)	<u>5.10</u>	<u>6.99</u>	<u>9.71</u>	<u>14.01</u>

Table 2: fwSNRseg results [dB] for baseline IVA and extended IVA (oracle) in four STFT frame length [samples] cases. The best fwSNRseg value in each case is underlined.

	512	1024	2048	8192
Baseline IVA	16.03	16.32	16.52	15.27
Extended IVA (oracle)	<u>19.29</u>	<u>20.52</u>	<u>22.02</u>	<u>24.24</u>

7.2 Speech and noise separation

Next, the results of the speech and noise separation experiments on extended IVA are presented; first, the results for the experiments on extended IVA without post-processing and then the results for the experiments involving post-processing.

7.2.1 Without post-processing

As it was explained in Section 6.2.1, source estimates used to compute the source model variances for extended IVA were obtained with two different speech enhancement methods: REPET SIM (see Section 6.2.2) and spectral subtraction (see Section 6.2.3). For clarity, these variants of extended IVA will be called from now on

extended IVA with REPET SIM and extended IVA with spectral subtraction, respectively. Also, we evaluated extended IVA with oracle variances as a reference of upper-bound performance. These variants of the extended IVA method were evaluated with four STFT window lengths in IVA, the same as for the speech separation experiment: 512, 1024, 2048 and 8192 samples. In each experiment, results were averaged over each of the three noise conditions of the data set: (a) cafeterias, (b) squares and (c) a subway car. The average result over the three noise environments is also shown. In these experiments the main comparison is between extended IVA with REPET SIM (or spectral subtraction) and baseline IVA. However, extended IVA with oracle variances, REPET SIM (or spectral subtraction) and the unprocessed observed mixtures were also evaluated for reference.

Figures 5 and 6 show the results for extended IVA with REPET SIM case. The results indicate that in general extended IVA with REPET SIM performed better than baseline IVA. For noise (a), SDR_i metric suggests that extended IVA with REPET SIM, with optimal frame length of 8192, performed nearly as good as REPET SIM, which was the best method for this noise condition. However, based on fwSNRseg, baseline IVA performs the best, with optimal frame length of 512 samples; the next best with a small margin of difference was extended IVA with REPET SIM (2048 optimal frame length). In case of noise (b), according to the SDR_i metric, the best method is extended IVA with REPET SIM (512 frame length); fwSNRseg in contrast favoured REPET SIM. For noise condition (c), extended IVA with REPET SIM is the method with best performance according to both metrics (at 1024 frame length for SDR_i, and 512 for fwSNRseg). On average, extended IVA with REPET SIM performed the best for both SDR_i and fwSNRseg (with 1024 and 512 frame lengths, respectively).

SDR_i and fwSNRseg results for extended IVA with spectral subtraction are presented in Figures 7 and 8, respectively. Extended IVA with spectral subtraction performs better than baseline IVA for all noise conditions for both SDR_i and fwSNRseg metrics. For noise (a), extended IVA with spectral subtraction gives the best performance, at 512 frame length for SDR_i and 1024 for fwSNRseg. For noise (b) extended IVA with spectral subtraction (at 512 frame length) achieves the best performance for SDR_i metric, while fwSNRseg indicates that spectral subtraction is superior in performance. In case of noise condition (c), extended IVA has the best performance, with 8192 frame length for SDR_i and 2048 frame length for fwSNRseg. On average, extended IVA with spectral subtraction is also the best method in both measures SDR_i and fwSNRseg (with frame lengths 512 and 2048, respectively).

On average, fwSNRseg measure favours extended IVA with REPET SIM (or extended IVA with spectral subtraction) over REPET SIM (or spectral subtraction) more than SDR_i does. Baseline IVA presents also better performance with respect to the other methods with fwSNRseg; for SDR_i, baseline IVA's performance is similar to that of the unprocessed observed mixtures. The difference on results obtained with the two measures is due to the different aspects of the signal each metric emphasizes. Listening to the audio samples indicated that REPET SIM and spectral subtraction remove more background noise than their extended IVA counterparts. Nonetheless, these two techniques introduce audible distortion in the speech signal, REPET SIM

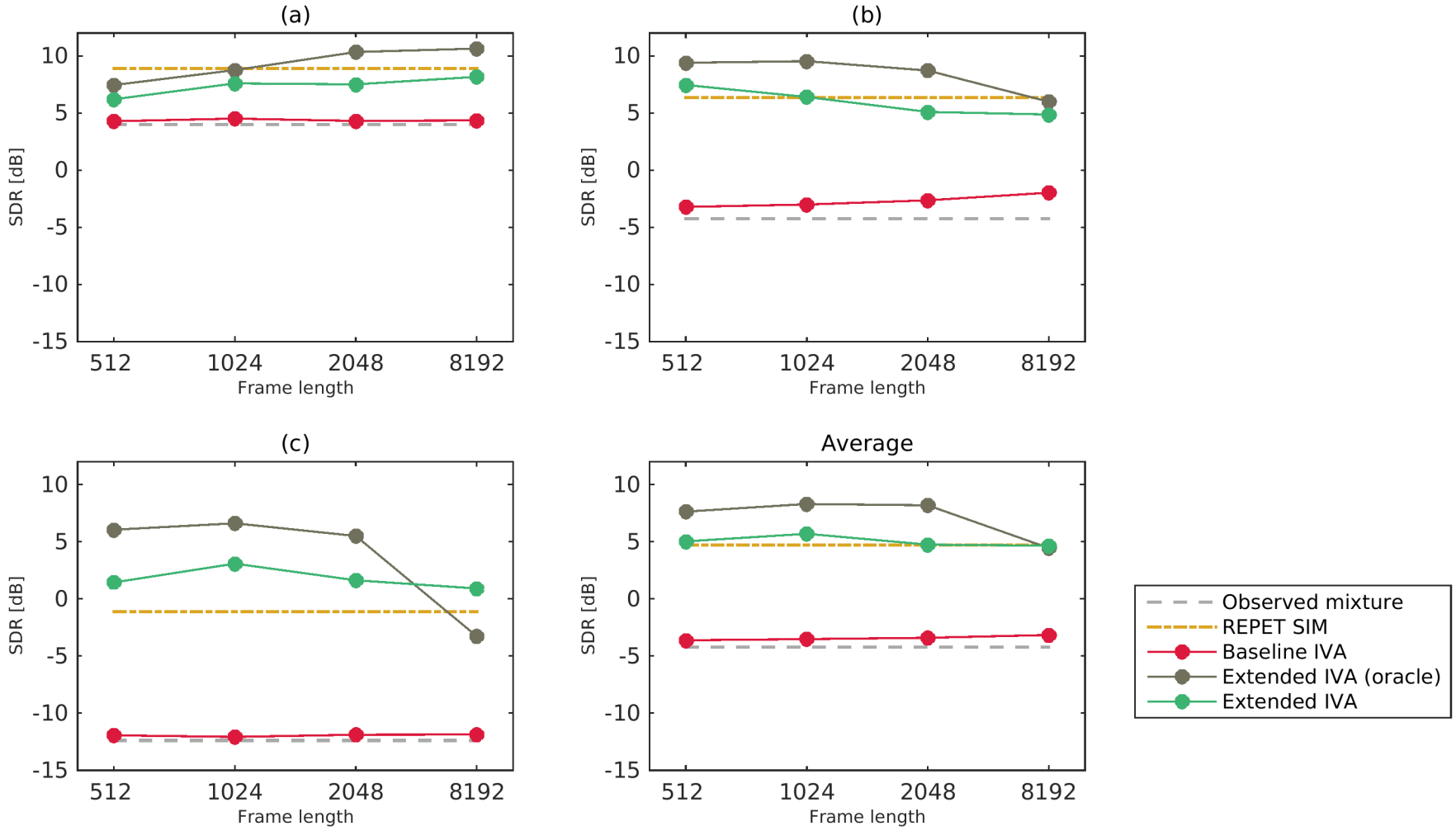

Figure 5: SDR_i results of the observed mixtures, REPET SIM and extended IVA with oracle variances (for reference), and the methods under comparison: baseline IVA and extended IVA with REPET SIM. Results for three street noise environments: (a) cafeterias, (b) squares and (c) a subway car, and the average over the three.

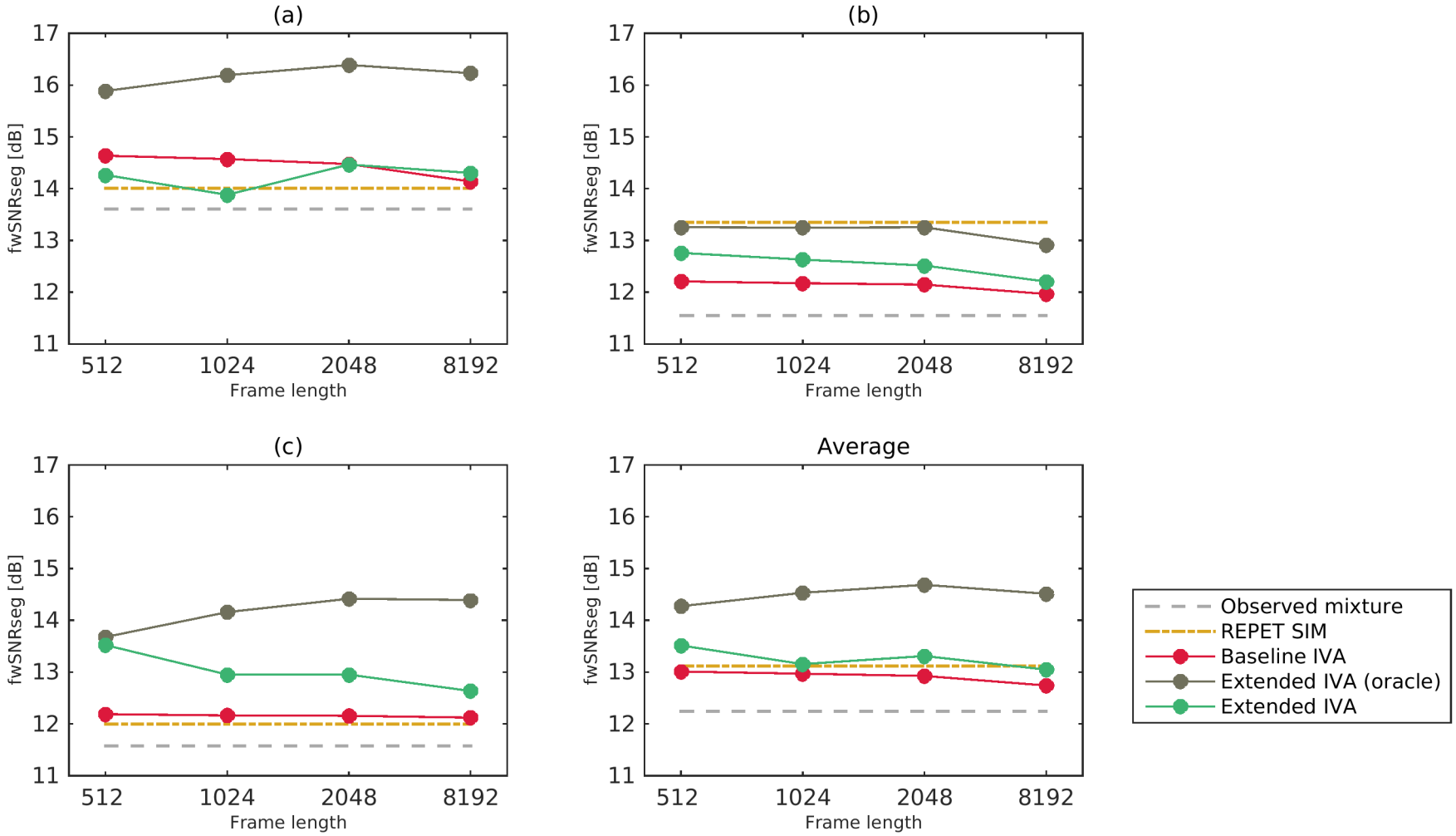

Figure 6: fwSNRseg results of the observed mixtures, REPET SIM and extended IVA with oracle variances (for reference), and the methods under comparison: baseline IVA and extended IVA with REPET SIM. Results for three street noise environments: (a) cafeterias, (b) squares and (c) a subway car, and the average over the three.

being the one that adds more distortion.

Finally, we address some details that stand out from the results presented. First, there is a clear difference between noise condition (b) (noise in public squares) and the other noise conditions (noise in cafeterias and noise in a subway car), for fwSNRseg measure. For noise (b), both REPET SIM and spectral subtraction have much better performance than their extended IVA counterparts and they are even better than the ideal case of extended IVA with oracle variances; this does not occur in the other noise cases. Second, in theory extended IVA with oracle variances should obtain the upper limit performance for extended IVA method. However, our results present several cases that do not follow this behaviour: in noise (c) and 8192 frame length, SDR_i metric indicates that both extended IVA with REPET SIM and extended IVA with spectral subtraction perform better than extended IVA with oracle variances; also, in noise (a) at 512 frame length, the SDR_i result from extended IVA with spectral subtraction is better than the result from the oracle version of extended IVA. Other detail we must point out is that, according to what we explained on frame length selection in Section 2.4.2, there should be a unique optimal frame length for extended IVA in each noise condition. However, we have seen that the results do not follow this behaviour: we have a different optimal frame length for each extended IVA variant (with REPET SIM, with spectral subtraction and with oracle variances). Finally, we note that optimal frame lengths also differ between the two metrics for the same noise condition, when actually we should have just one optimal frame length per method and noise condition. This could be explained by the fact that each metric emphasizes different qualities of the signal. In any case, even though we pick here an optimal frame length, in general, the results for each method do not fluctuate much within frame lengths 512-8192.

7.2.2 With post-processing

The following experiments, involving post-processing, were conducted on only one of the two extended IVA variants compared to baseline IVA in the experiments of Section 7.2.1. Extended IVA with spectral subtraction was chosen over extended IVA with REPET SIM based on informal listening of the audio samples that indicated that spectral subtraction was perceptually better than REPET SIM.

We performed experiments in which extended IVA was evaluated with several post-filters and with spectral subtraction post-filtering. The post-processing aimed to reduce the noise present in the separated speech obtained with extended IVA to further improve the source separation performance. Details on our post-processing approach and the post-filters employed are given in Section 5.2. As in the experiments without post-processing (Section 7.2.1), these experiments were carried out with four different frame lengths of IVA: 512, 1024, 2048 and 8192 samples. Here also the results were averaged over each of the three noise conditions in the data set, and the average result over these three noises was also computed.

The results of extended IVA with spectral subtraction without and with post-processing from each of the three post-filters and spectral subtraction technique are shown in Figures 9 and 10. Results on the SDR_i and fwSNRseg metrics indicate

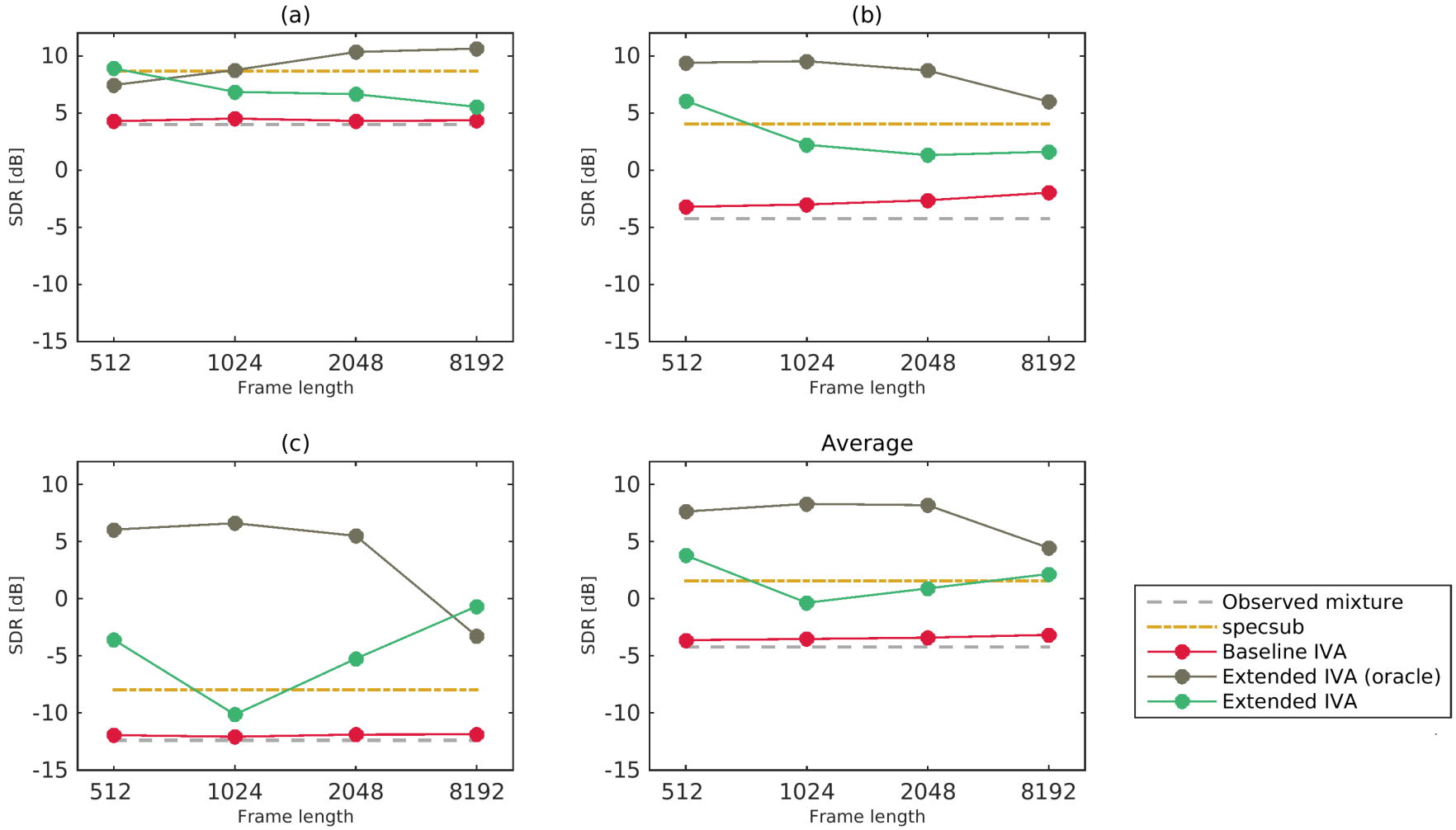

Figure 7: SDR results of the observed mixtures, spectral subtraction and extended IVA with oracle variances (for reference), and the methods under comparison: baseline IVA and extended IVA with spectral subtraction. Results for three street noise environments: (a) cafeterias, (b) squares and (c) a subway car, and the average over the three.

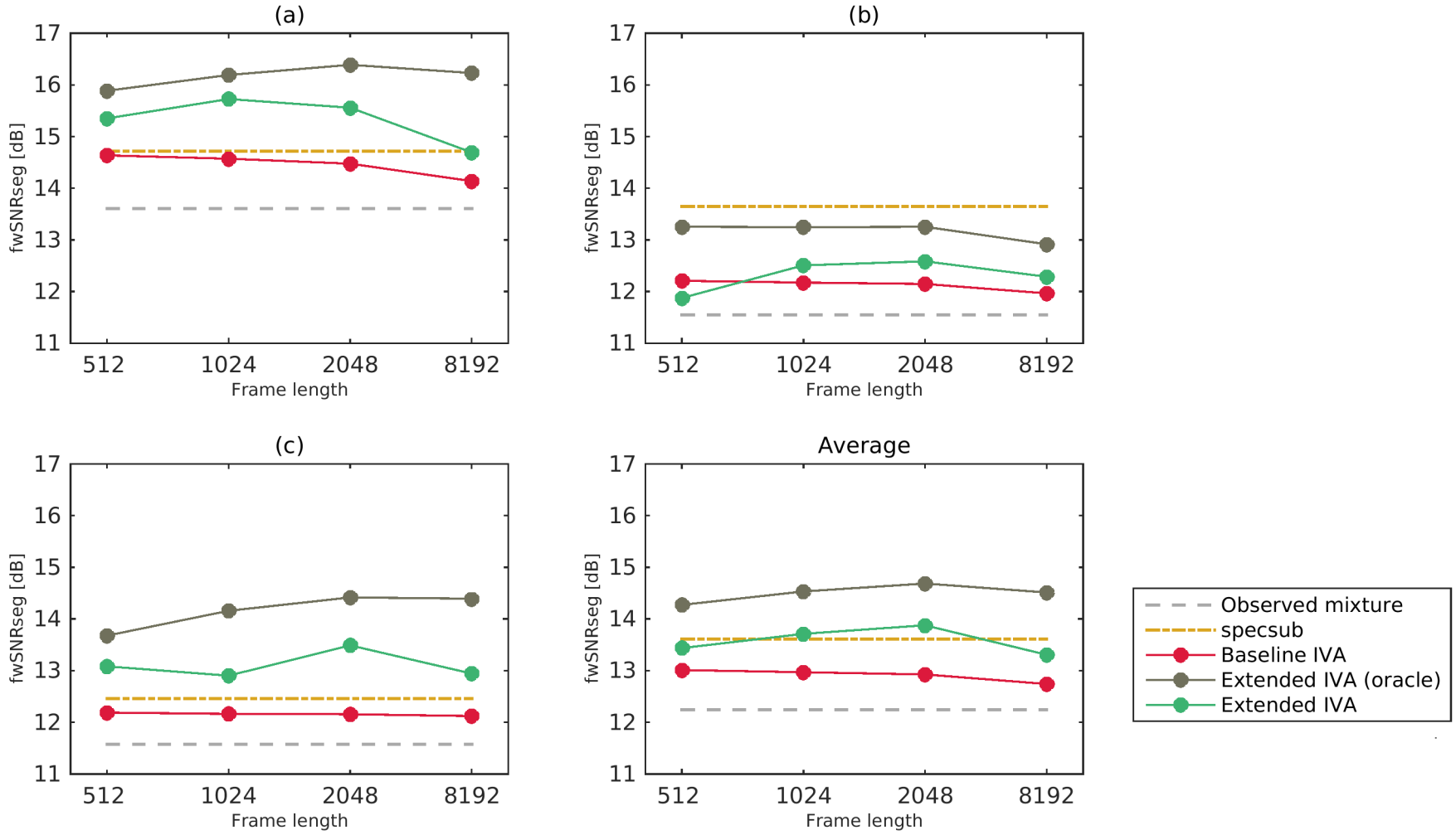

Figure 8: fwSNRseg results of the observed mixtures, spectral subtraction and extended IVA with oracle variances (for reference), and the methods under comparison: baseline IVA and extended IVA with spectral subtraction. Results for three street noise environments: (a) cafeterias, (b) squares and (c) a subway car, and the average over the three.

that extended IVA is generally better with post-processing than without. In noise condition (a), the results of the different post-filtering solutions are very tight for SDR_i metric; the best performance is obtained using extended IVA with $H^{(1)}$ post-filter and extended IVA with spectral subtraction post-filtering (both at frame length 512), and $H^{(2)}$ and $H^{(3)}$ post-filters are almost as good. In case of fwSNRseg metric, both extended IVA without post-filtering and with spectral subtraction post-filtering give the best results (with frame length 1024). Therefore in this case, except for spectral subtraction post-filtering, extended IVA gives better results without post-processing. For noise (b), extended IVA with $H^{(1)}$ post-filter and extended IVA with spectral subtraction post-filtering show the best results, with frame length 512, according to SDR_i metric. In case of fwSNRseg, all the post-filtering solutions improve the performance of extended IVA with a clear margin. Among these post-filtering solutions, spectral subtraction post-filtering is the one that gives the best results (with frame length 8192). In case of noise (c), the best result is obtained for extended IVA with $H^{(1)}$, at frame length 8192 for SDR_i metric. For fwSNRseg, the best result is obtained also for extended IVA with $H^{(1)}$ (at frame length 2048). On average, the best result with SDR_i measure is obtained with spectral subtraction post-filtering and $H^{(1)}$ post-filter (at frame length 512). The SDR_i results also indicate that on average $H^{(3)}$ performs worse than $H^{(1)}$ and $H^{(2)}$. This suggests that the two Wiener filters $H^{(1)}$ and $H^{(2)}$ are in general more efficient at noise reduction than $H^{(3)}$, the amplitude replacing filter. Besides, listening to the audio samples support this. In case of fwSNRseg, the best method on average is extended IVA with spectral subtraction post-filtering, for frame length 2048.

Figures 11 and 12 show a comparison of results from the best post-filter solution at each noise case against the results from the observed mixtures, spectral subtraction technique, baseline IVA and extended IVA; $H^{(1)}$ was selected the best post-filter for all noise conditions. We excluded from the selection spectral subtraction post-filtering, since it is computationally much more complex than the post-filter solutions and does not present substantially better performance. In addition, we should note that the selection of best post-filter in each case is not definitive; as we stated before, in many cases the results between several post-filtering solutions were very similar, which suggests that the differences between these post-filters are not statistically significant.

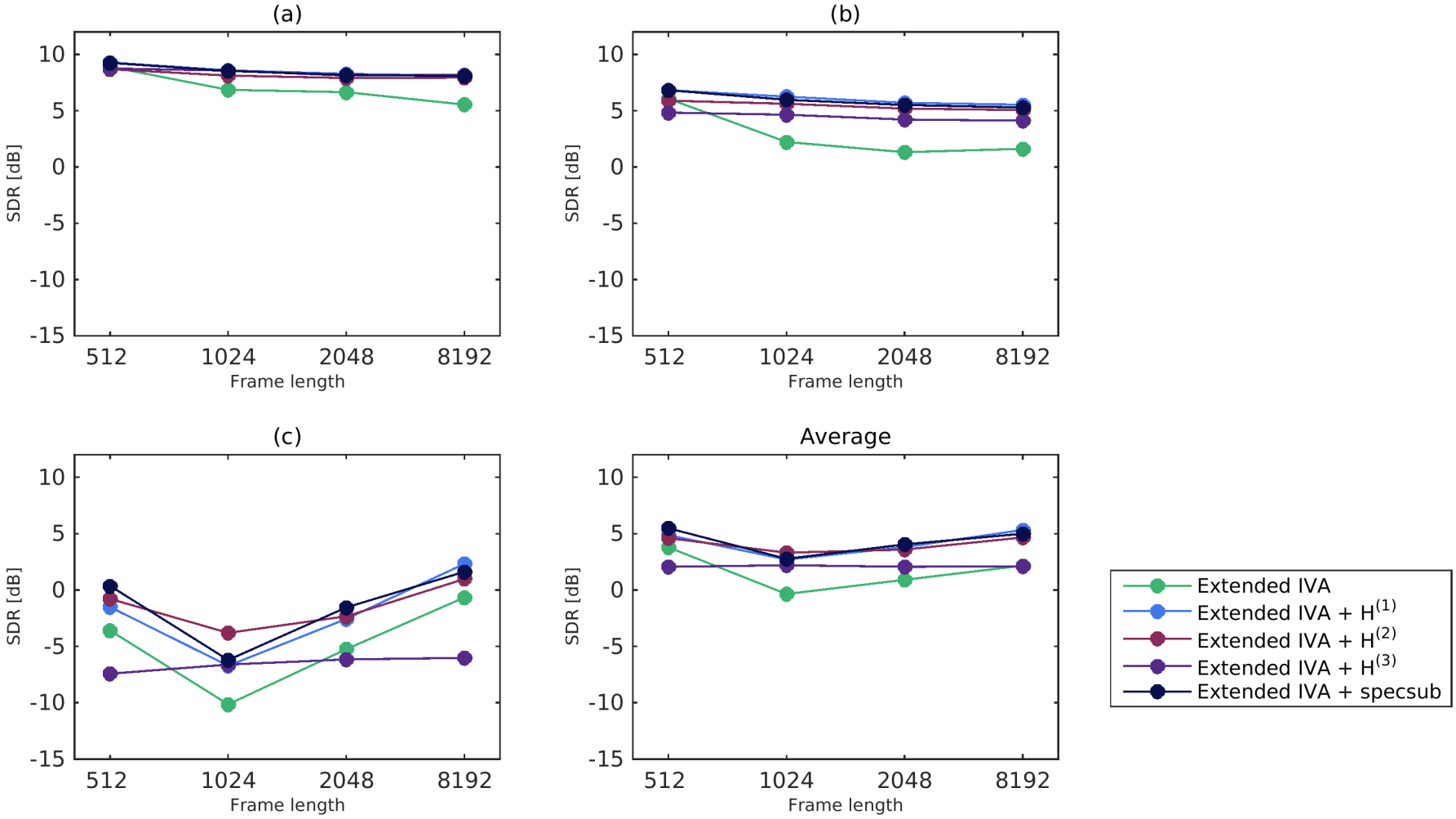

Figure 9: SDR_i results of extended IVA (with spectral subtraction), and extended IVA (with spectral subtraction) with post-filters $H^{(1)}$, $H^{(2)}$ and $H^{(3)}$; and extended IVA (with spectral subtraction) with spectral subtraction post-filtering. Results for three street noise environments: (a) cafeterias, (b) squares and (c) a subway car, and the average over the three.

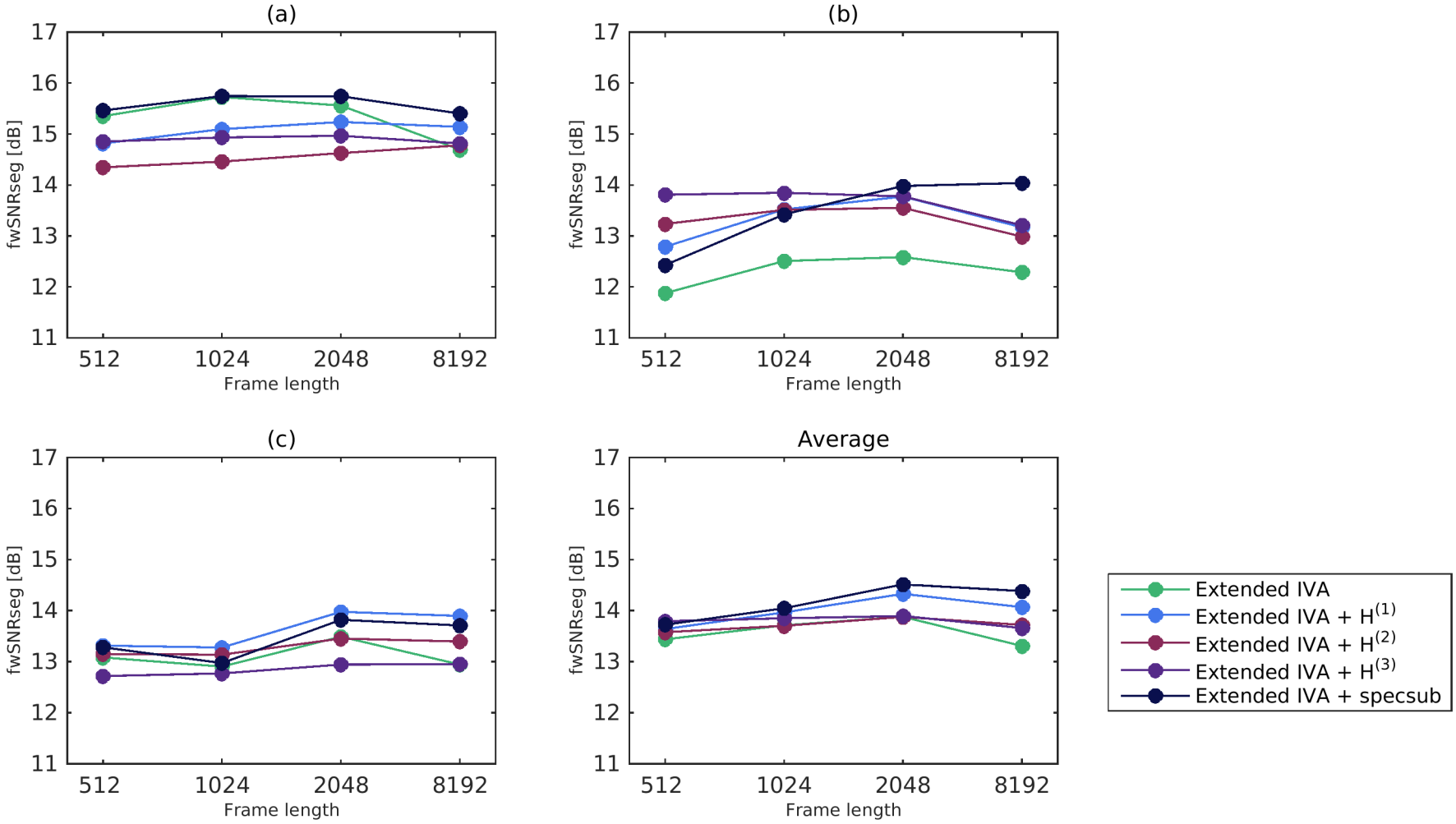

Figure 10: fwSNRseg results of extended IVA (with spectral subtraction), and extended IVA (with spectral subtraction) with post-filters $H^{(1)}$, $H^{(2)}$ and $H^{(3)}$; and extended IVA (with spectral subtraction) with spectral subtraction post-filtering. Results for three street noise environments: (a) cafeterias, (b) squares and (c) a subway car, and the average over the three.

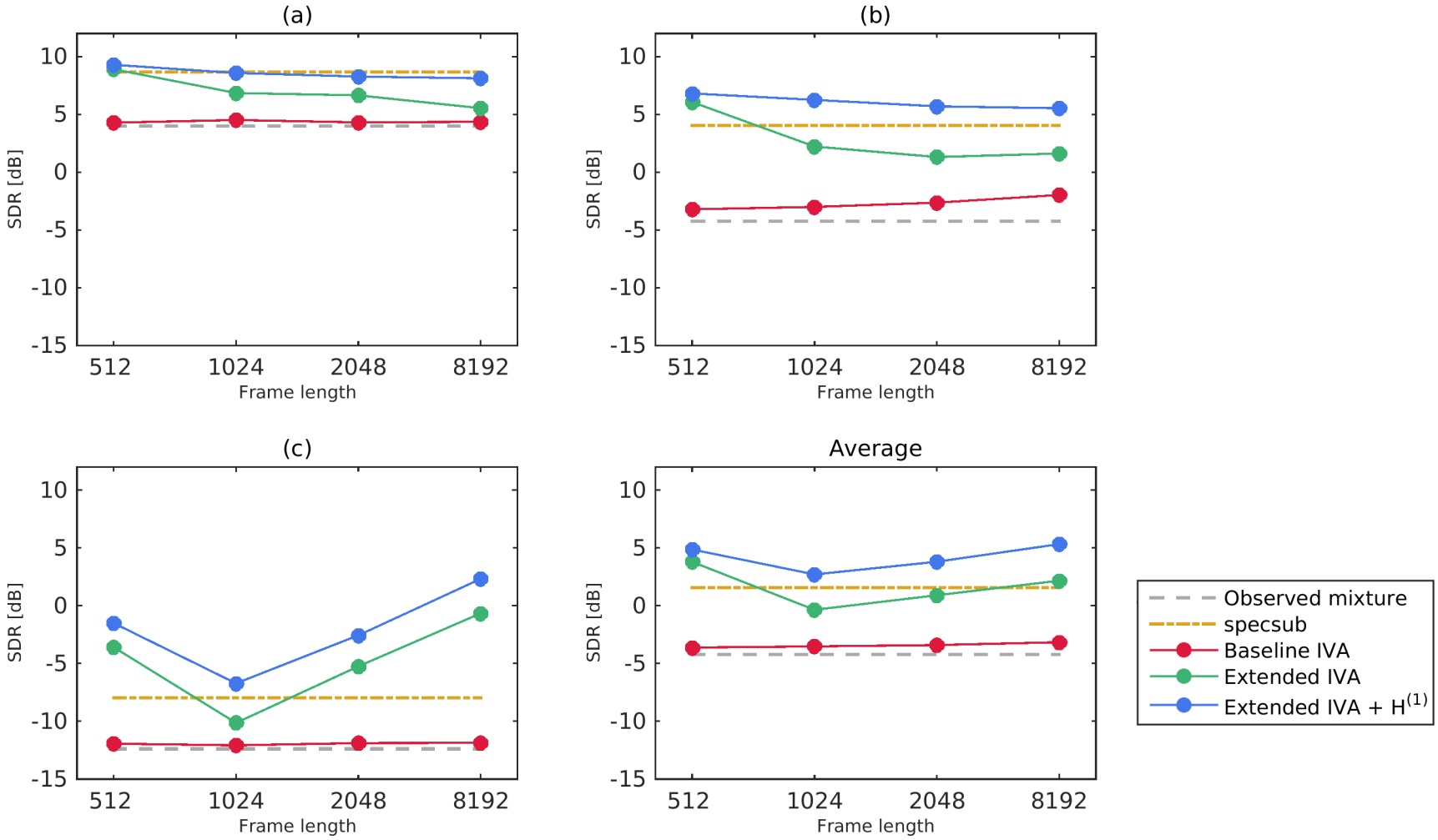

Figure 11: SDR_i results of extended IVA (with spectral subtraction) without and with post-filter (best post-filter for each case); and observed mixtures, spectral subtraction technique and baseline IVA for reference. Results for three street noise environments: (a) cafeterias, (b) squares and (c) a subway car, and the average over the three.

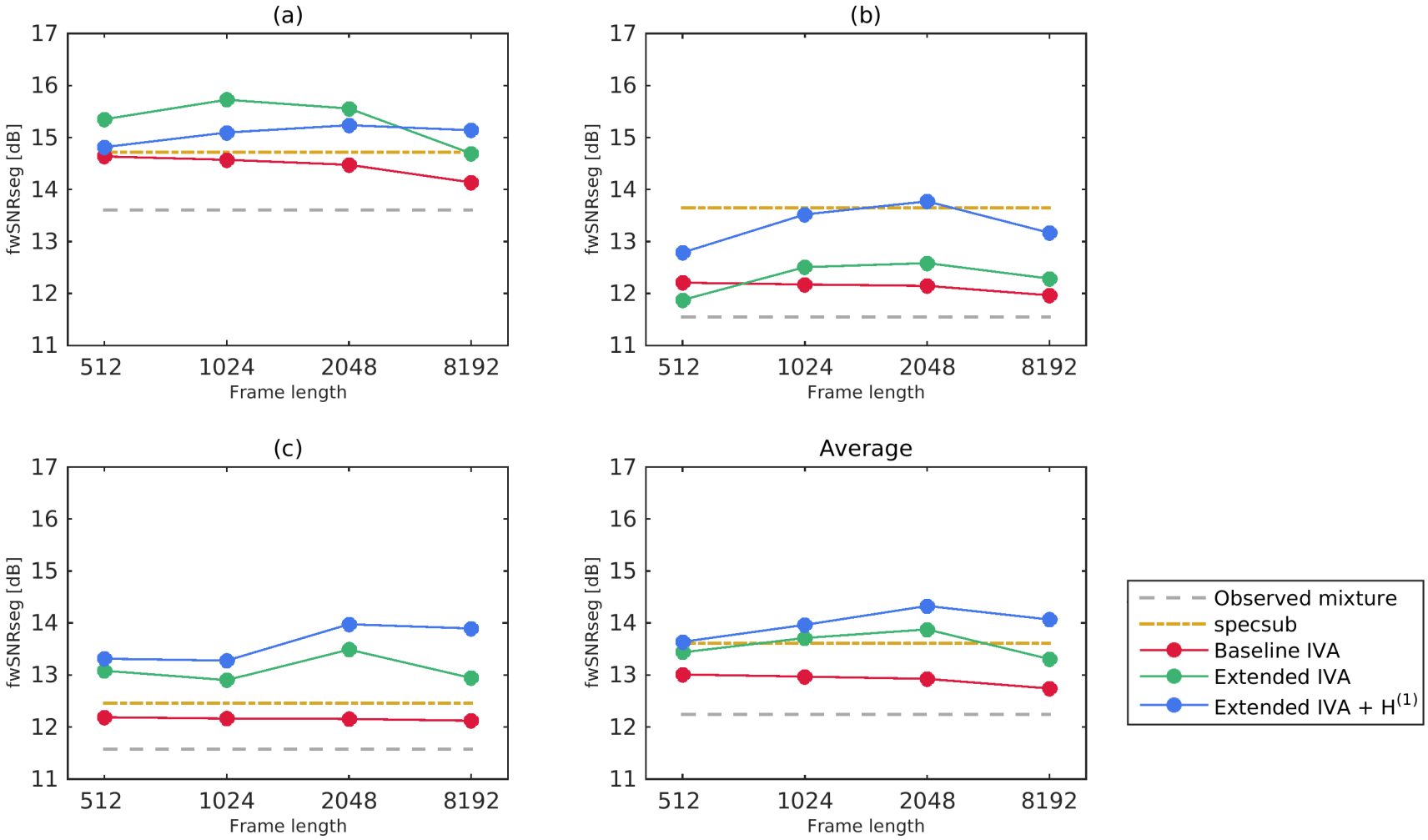

Figure 12: fwSNRseg results of extended IVA (with spectral subtraction) without and with post-filter (best post-filter for each case); and observed mixtures, spectral subtraction and baseline IVA for reference. Results for three street noise environments: (a) cafeterias, (b) squares and (c) a subway car, and the average over the three.

8 Discussion

Three main points of discussion are presented in this section, based on the goals we had and which were introduced in Section 1: first, the improvement of IVA's performance by using new, improved source models for speech and noise separation (Section 8.1); second, to diminish the diffuse noise reduction limitation of IVA with post-processing solutions (Section 8.2); and third, the selection of suitable measures for evaluation of the method proposed (Section 8.3). In addition, we discuss in Section 8.4 about possible future directions of our research.

8.1 Improving IVA with new source models

In the present work, we extended IVA methods with a new, improved source model for speech and noise separation. The proposed approach of IVA is referred as extended IVA, and the source model employed is a time-frequency-variant Gaussian distribution. With this model, we aimed to overcome the limitations for speech and noise separation tasks of traditional source models used in IVA, and as a result improve the source separation performance of IVA. Even though this work focuses on IVA for speech and noise separation (specifically, diffuse noise), we also ran a preliminary experiment on a speech separation task, where the signals to be separated consisted of mixtures of two speech signals.

A signal originating from a speaker is considered a *localized source*, that is, a source that originates from an specific direction. Localized sources are the optimal source signals that IVA can handle, since IVA is basically performing spatial filtering: it filters the arriving signals according to the location on the space. In contrast to localized sources, diffuse noise originates from many directions, and therefore it is a difficult task for IVA. What we pretended with the speech separation experiment was to validate the extended IVA approach first under the best conditions for IVA, with localized sources. If the results from this experiment would have not proved successful, there would have not been reason to continue further with the approach proposed. In addition, in the speech separation task, extended IVA used oracle source models, meaning that the true sources were employed for computing the source model. We did this to test extended IVA without dependency on the performance of the support method used to compute the source variances. In other words, we avoid the situation in which extended IVA might fail because of the support method but not because of the extended IVA approach itself. Once this experiment was performed with success, we could evaluate extended IVA under more challenging conditions, in which speech had to be separated from diffuse noise. We evaluated extended IVA first using oracle source models that showed the upper limit performance of the method for this condition. Finally, extended IVA was evaluated in a more realistic situation, using source variances computed with a support method.

As we said, extended IVA was evaluated with two tasks: a speech separation task with two-channel mixtures of two speech signals and a speech and noise separation task with two-channel mixtures of speech and diffuse noise. Comparing the results of

extended IVA and baseline IVA from both tasks validated our hypothesis that IVA source separation performance improves when improved source models are employed. Our results are consistent with the work presented in [39]: the proposed IVA method, which uses a time-varying source model, performs better than the baseline IVA method used for comparison. In addition to this, our results for the speech and noise separation task can be directly compared to the SISEC 2013's results[6] of the same task and data set; three methods were evaluated for this task in SISEC 2013. The comparison is based on SDR_i results, since this is the common metric of both works. The results obtained with extended IVA (with REPET SIM or spectral subtraction), at optimal frame length, are comparable to those from the methods evaluated in SISEC 2013; in average, the performance of extended IVA is close to the performance of the SISEC 2013's method ranked as second best. Nevertheless, the performance of extended IVA with oracle variances, at optimal frame length, is close to the performance of the best method; this shows clearly the potential of extended IVA.

Finally, we comment on the significance of the results of the tasks in this work. The datasets of the speech separation and speech and noise separation tasks consisted of 48 and 9 mixtures, respectively. Given the limited amount of samples tested, specially in the speech and noise separation task, one may conclude that the results obtained cannot be generalized to other noise conditions.

8.2 Post-processing for diffuse noise reduction

A second goal of this work was to further improve the performance of extended IVA. For that, we applied post-filtering to the speech outputs of extended IVA. We hypothesized that the post-filtering would reduce the diffuse noise that IVA is not capable of removing because of the multi-directional nature of this kind of source signal. Our hypothesis is based on previous research in which applying post-filtering to a beamformer has proven to reduce the diffuse noise remaining on the output signals of the beamformer [15]. We evaluated extended IVA with three different post-filter solutions, and also with spectral subtraction post-filtering. The results indicated that the performance of extended IVA improved with post-filtering. Informal listening of the audio samples indicates there is a noise reduction with the post-filtering, as expected, at the cost of some distortion added to the samples. This is reflected also in the SDR_i and fwSNRseg results, since the post-filtering improvement is more prominent with the SDR_i metric. In conclusion, these results confirm our hypothesis. Besides, they are in line with previous works on the topic [15, 47], in which applying post-filtering after the multichannel source separation system increases the noise reduction from the source estimates. The improvement obtained by applying post-processing after extended IVA is also reflected when comparing our results with SISEC 2013's results for the same task; as we discussed in Section 8.1, the results from extended IVA without post-processing are in average close to the results obtained with the second best ranked method in SISEC 2013, while the

[6] http://www.onn.nii.ac.jp/sisec13/evaluation_result/BGN/homepage_BGN_dev.html

performance of extended IVA with post-processing is in general superior to the performance of the second best method.

Out of the three post-filters, the Wiener post-filter $H^{(1)}$ showed the best performance. Compared to $H^{(1)}$ performance, spectral subtraction post-filtering got similar results for SDR_i metric and performed better for fwSNRseg. However, in extended IVA with spectral subtraction post-filtering we have to compute spectral subtraction in two different stages: first, spectral subtraction's source estimates are computed for the source model of extended IVA, and then when spectral subtraction is computed for post-processing. In contrast, extended IVA with post-filter $H^{(1)}$, $H^{(2)}$ or $H^{(3)}$ only needs to compute spectral subtraction in the first step, but the post-filter does not involve spectral subtraction. Therefore, the post-filters have the advantage that they are computationally less complex.

8.3 Evaluation measures

The last goal of this work was to select suitable measures to evaluate accurately the performance of the method proposed. We started our work with one of the standard metrics for source separation: an energy ratio, SDR_i [59, 60]. Some early experiments evaluated with SDR_i, compared to audio listening of the actual samples, made us realize that this metric did not reflect much about the perceptual quality and focused more on the interference present in the enhanced signal. As a result, we decided to use also another measure that would emphasize more perceptual aspects of the signals, and it could complement the evaluation with SDR_i. We chose fwSNRseg as a suitable measure because of its high correlation to subjective listening tests [61].

The results of the experiments with these two metrics, matched with informal audio listening, proved that using both measures was important for the proper evaluation of the method. Each one of the two metrics focused on different aspects of the signal. In consequence, the method that was the best for one metric may not be also the best for the other. That was the case, for example, when comparing spectral subtraction and extended IVA: on average SDR_i favoured spectral subtraction over extended IVA, but in case of fwSNRseg extended IVA obtained better results.

8.4 Future work

In the present work, we obtained improvements with extended IVA over baseline IVA; even when the source models are provided by simple support methods like the ones used, REPET SIM and spectral subtraction. However, it must be noted that even though in our work the source variances were calculated as Equation (19), based on source estimates from a support method, they do not need to be determined in this manner. In fact, since the source variances are computed as proposed in Equation (19), extended IVA can be used in speech separation tasks only for the oracle case in which the true sources are known. This is because the support methods that we employ are only able to separate a target signal from a background signal. Therefore, other approach for computing the source variances could be used; one

that does not require obtaining source estimates from a support method and as a result, extended IVA could be employed in speech separation without needing oracle information.

Other possible next step for research in extended IVA could be using other support methods that are more sophisticated than REPET SIM or spectral subtraction. For example, methods like NMF [65] or deep neural networks (DNN) [66] could be good candidates for obtaining better source estimates. This should lead to better results of extended IVA, since our results of extended IVA with oracle variances have proved that there is still margin for improving the performance of extended IVA. With regards to the current support methods employed, it must be mentioned that extended IVA with spectral subtraction was also evaluated on CHiME dataset [67], and the results of extended IVA did not turn out to be successful. CHiME dataset consists of noisy speech and the noise component is often non-stationary. Spectral subtraction technique works under the assumption of stationary noise, and therefore it fails to give satisfactory source estimates for non-stationary noise. Therefore, this suggests that the negative outcome obtained in this experiment was a result of the bad performance of the support method and not of extended IVA; reinforcing the idea that using other support methods would be beneficial for extended IVA, since they could also make extended IVA more versatile if they get good estimates for varying real-world noisy conditions.

References

[1] CHABA (Committee on Hearing, Bioacoustics, and Biomechanics), "Speech understanding and aging," *Journal of the Acoustical Society of America*, vol. 83, no. 3, pp. 859–895, 1988.

[2] J. Cunningham, T. Nicol, S. G. Zecker, A. Bradlow, and N. Kraus, "Neurobiologic responses to speech in noise in children with learning problems: deficits and strategies for improvement," *Clinical Neurophysiology*, vol. 112, no. 5, pp. 758–767, 2001.

[3] R. J. Turner, "Social support as a contingency in psychological well-being," *Journal of Health and Social Behavior*, vol. 22, no. 4, pp. 357–367, 1981.

[4] P. C. Loizou, *Speech Enhancement: Theory and Practice.* CRC Press, 2013.

[5] J. Benesty, S. Makino, and J. Chen, Eds., *Speech Enhancement.* Springer, 2005.

[6] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[7] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustica united with Acustica*, vol. 86, no. 1, pp. 117–128, 2000.

[8] Y. A. Huang, J. Benesty, and J. Chen, "Separation and dereverberation of speech signals with multiple microphones," in *Speech Enhancement.* Springer, 2005, pp. 271–298.

[9] S. C. Douglas, "Blind separation of acoustic signals," in *Microphone Arrays.* Springer, 2001, pp. 355–380.

[10] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[11] Y. Deville, C. Jutten, and R. Vigario, "Overview of source separation applications," in *Handbook of Blind Source Separation.* Academic Press, 2010, pp. 639–681.

[12] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *Proc. ICA*, 2006, pp. 165–172.

[13] A. Hiroe, "Solution of Permutation Problem in Frequency Domain ICA, Using Multivariate Probability Density Functions," in *Proc. ICA*, 2006, pp. 601–608.

[14] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 70–79, 2007.

[15] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays*. Springer, 2001, pp. 39–60.

[16] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. John Wiley & Sons, 2001.

[17] T.-W. Lee, *Independent Component Analysis: Theory and Applications*. Kluwer Academic Publishers, 1998.

[18] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[19] A. Cichocki, R. Zdunek, and S. Amari, "New algorithms for non-negative matrix factorization in applications to blind source separation," in *Proc. ICASSP*, 2006, pp. 621–624.

[20] R. Gribonval and M. Zibulevsky, "Sparse component analysis," in *Handbook of Blind Source Separation*. Academic Press, 2010, pp. 367–420.

[21] Y. Li, A. Cichocki, and S. Amari, "Analysis of sparse representation and blind source separation," *Neural Computation*, vol. 16, pp. 1193–1234, 2004.

[22] S. C. Douglas and M. Gupta, "Convolutive blind source separation for audio signals," in *Blind Speech Separation*. Springer, 2007, pp. 3–46.

[23] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, "A survey of convolutive blind source separation methods," in *Springer Handbook of Speech Processing*. Springer, 2007.

[24] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Frequency-domain blind source separation," in *Speech Enhancement*. Springer, 2005, pp. 299–327.

[25] H.-L. N. Thi and C. Jutten, "Blind source separation for convolutive mixtures," *Signal Processing*, vol. 45, no. 2, pp. 209–229, 1995.

[26] A. Hyvärinen, "Survey on independent component analysis," *Neural Computing Surveys*, vol. 2, pp. 94–128, 1999.

[27] J. F. Cardoso, "Multidimensional independent component analysis," in *Proc. ICASSP*, 1998, pp. 1941–1944.

[28] A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Computation*, vol. 9, no. 7, pp. 1483–1492, 1997.

[29] J. F. Cardoso, "Infomax and maximum likelihood for blind source separation," *IEEE Letters on Signal Processing*, vol. 4, no. 4, pp. 112–114, 1997.

[30] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[31] J. F. Cardoso and A. Souloumiac, "Blind beamforming for non-Gaussian signals," in *Proc. F (Radar and Signal Processing)*, vol. 140, no. 6, 1993, pp. 362–370.

[32] J. F. Cardoso, "High-order contrasts for independent component analysis," *Neural Computation*, vol. 11, no. 1, pp. 157–192, 1999.

[33] S. Makino, H. Sawada, R. Mukai, and S. Araki, "Blind source separation of convolutive mixtures of speech in frequency domain," *IEICE Trans. Fundamentals of Electronics, Communications and Computer Sciences*, vol. 88, no. 7, pp. 1640–1655, 2005.

[34] T. Nishikawa, H. Saruwatari, and K. Shikano, "Blind source separation of acoustic signals based on multistage ICA combining frequency-domain ICA and time-domain ICA," *IEICE Trans. Fundamentals of Electronics, Communications and Computer Sciences*, vol. 86, no. 4, pp. 846–858, 2003.

[35] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 3, pp. 320–327, 2000.

[36] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 2, pp. 109–116, 2003.

[37] M. Z. Ikram and D. R. Morgan, "Permutation inconsistency in blind speech separation: investigation and solutions," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 1, pp. 1–13, 2005.

[38] S. Roberts and R. Everson, *Independent Component Analysis: Principles and Practice*. Cambridge University Press, 2001.

[39] T. Ono, N. Ono, and S. Sagayama, "User-guided independent vector analysis with source activity tuning," in *Proc. ICASSP*, 2012, pp. 2417–2420.

[40] N. Ono, "Auxiliary-function-based independent vector analysis with power of vector-norm type weighting functions," in *Proc. APSIPA*, 2012.

[41] ——, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. WASPAA*, 2011, pp. 189–192.

[42] N. Ono and S. Miyabe, "Auxiliary-function-based independent component analysis for super-Gaussian sources," in *Proc. LVA/ICA*, 2010, pp. 165–172.

[43] N. Ono, "Fast stereo independent vector analysis and its implementation on mobile phone," in *Proc. IWAENC*, 2012, pp. 1–4.

[44] Z. Rafii and B. Pardo, "Online REPET-SIM for real-time speech enhancement," in *Proc. ICASSP*, 2013, pp. 848–852.

[45] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. ICASSP*, vol. 4, 1979, pp. 208–211.

[46] J. Li and M. Akagi, "A noise reduction system based on hybrid noise estimation technique and post-filtering in arbitrary noise environments," *Speech Communication*, vol. 48, no. 2, pp. 111–126, 2006.

[47] I. A. McCowan and H. Bourlard, "Microphone array post-filter for diffuse noise field," in *Proc. ICASSP*, vol. 1, 2002, pp. 905–908.

[48] S. Gannot and I. Cohen, "Adaptive beamforming and postfiltering," in *Springer Handbook of Speech Processing*. Springer, 2008, pp. 945–978.

[49] S. Araki, Y. Hinamoto, S. Makino, T. Nishikawa, R. Mukai, and H. Saruwatari, "Equivalence between frequency domain blind source separation and frequency domain adaptive beamforming," in *Proc. ICASSP*, vol. 2, 2002, pp. 1785–1788.

[50] C. Marro, Y. Mahieux, and K. U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 3, pp. 240–259, 1998.

[51] S. V. Vaseghi, "Wiener filters," in *Advanced Digital Signal Processing and Noise Reduction*. John Wiley & Sons, 2001, pp. 178–204.

[52] R. M. Parry and I. Essa, "Incorporating phase information for source separation via spectrogram factorization," in *Proc. ICASSP*, vol. 2, 2007, pp. 661–664.

[53] E. Vincent, S. Araki, and P. Bofill, "The 2008 Signal Separation Evaluation Campaign: A community-based approach to large-scale evaluation," in *Proc. ICA*, 2009, pp. 734–741.

[54] N. Ono, Z. Koldovsky, S. Miyabe, and N. Ito, "The 2013 Signal Separation Evaluation Campaign," in *Proc. MLSP*, 2013, pp. 1–6.

[55] M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," 2006, available at http://www.ee.imperial.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

[56] D. Schobben, K. Torkkola, and P. Smaragdis, "Evaluation of blind signal separation methods," in *Proc. ICA*, 1999, pp. 261–266.

[57] A. Mansour, M. Kawamoto, and N. Ohnishi, "A survey of the performance indexes of ICA algorithms," in *Proc. MIC*, 2002, pp. 660–666.

[58] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[59] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, "First stereo audio source separation evaluation campaign: Data, algorithms and results," in *Proc. ICA*, 2007, pp. 552–559.

[60] E. Vincent, S. Araki, F. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, V. Gowreesunker, D. Lutter, and N. Q. K. Duong, "The Signal Separation Evaluation Campaign (2007-2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, no. 8, pp. 1928–1936, 2012.

[61] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[62] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality*. Prentice-Hall, 1988.

[63] "Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," ITU-T Recommendation P.862, 2001.

[64] R. Karhila, U. Remes, and M. Kurimo, "Noise in HMM-based speech synthesis adaptation: Analysis, evaluation methods and experiments," *IEEE Journal of Selected Topics in Signal Processing*, 2014.

[65] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067–2080, 2011.

[66] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[67] H. Christensen, J. Barker, N. Ma, and P. Green, "The CHiME corpus: a resource and a challenge for computational hearing in multisource environments," in *Proc. INTERSPEECH*, 2010, pp. 1918–1921.